INTRODUCTION TO DATA ANALYTICS

Class #5

Descriptive Statistics II

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

TODAY'S DISCUSSION…

- Measurement of location
  - Mean, median, mode, midrange, etc.
- Measure of dispersion
  - Range, Variance, Standard Deviation, etc.
- Other measures
  - MAD, AAD, Percentile, IQR, etc.

- Graphical summarization
  - Box plot

AM, GM AND HM

- Is there any generalization for AM ($\bar{x}$), GM ($\tilde{x}$) and HM ($\hat{x}$) calculations for a sample of size ≥ 2?
- In which situation, a particular mean is applicable?
- If there is any interrelationship among them?

GEOMETRIC MEAN

Definition 3.9: **Geometric mean**

Geometric mean of $\boldsymbol{n}$ observations (*none of which are zero*) is defined as:

$$\tilde{x} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

where, $n \neq 0$

Note

- GM is the arithmetic mean in "log space". This is because, alternatively,

$$log\tilde{x} = \frac{1}{n} \sum_{i=1}^{n} log\, x_i$$

- This summary of measurement is meaningful only when all observations are > 0
  - If at least one observation is zero, the product will itself be zero! For a negative value, root is not real

HARMONIC MEAN

Definition 3.10: **Harmonic mean**

If all observations are non zero, the reciprocal of the arithmetic mean of the reciprocals of observations is known as harmonic mean.

For ungrouped data

$$\hat{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

For grouped data

$$\hat{x} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \left(\frac{f_i}{x_i}\right)}$$

where, f_i is the frequency of the i^{th} class with x_i as the center value of the i^{th} class.

SIGNIFICANT OF DIFFERENT MEAN CALCULATIONS

- There are two things involved when we consider a sample
  - Observation
  - Range

Example: Rainfall data

Rainfall (in mm)	r_1	r_2	…	r_n
Days (in number)	d_1	d_2	…	d_n

- Here, rainfall is the observation and day is the range for each element in the sample
- Here, we are to measure the mean "**rate of rainfall**" as the measure of location

SIGNIFICANT OF DIFFERENT MEAN CALCULATIONS

- **Case 1: Range remains same for each observation**

Example: Having data about amount of rainfall per week, say.

Rainfall (in mm)	35	18	…	22
Days (in number)	7	7	…	7

SIGNIFICANT OF DIFFERENT MEAN CALCULATIONS

- **Case 2: Ranges are different, but observation remains same**

Example: Same amount of rainfall in different number of days, say.

Rainfall (in mm)	50	50	…	50
Days (in number)	1	2	…	7

SIGNIFICANT OF DIFFERENT MEAN CALCULATIONS

- **Case 3: Ranges are different, as well as the observations**

Example: Different amount of rainfall in different number of days, say.

Rainfall (in mm)	21	34	…	18
Days (in number)	5	3	…	7

RULE OF THUMBS FOR MEANS

- **AM:** When the range remains same for each observation

  Example: Case 1

Rainfall (in mm)	35	18	…	22
Days (in number)	7	7	…	7

$$\bar{r} = \frac{1}{n}\sum_{1}^{n} r_i$$

RULE OF THUMBS FOR MEANS

- **HM:** When the range is different but each observation is same
  - Example: Case 2

Rainfall (in mm)	50	50	…	50
Days (in number)	1	2	…	7

$$\tilde{r} = \frac{n}{\sum_1^n \frac{1}{r_i}}$$

RULE OF THUMBS FOR MEANS

- **GM:** When the ranges are different as well as the observations
  - Example: Case 3

Rainfall (in mm)	21	34	…	18
Days (in number)	5	3	…	7

$$\hat{r} = \left(\prod_{1}^{n} r_i \right)^{\frac{1}{n}}$$

RULE OF THUMBS FOR MEANS

- The important things to recognize is that all three means are simply the arithmetic means in disguise!

- Each mean follows the "additive structure".

  - Suppose, we are given some abstract quantities $\{x_1, x_2, \ldots, x_n\}$

  - Each of the three means can be obtained with the following steps

    1. Transform each x_i into some y_i

    2. Taking the arithmetic mean of all y_i's

    3. Transforming back the to the original scale of measurement

RULE OF THUMBS FOR MEANS

- For arithmetic mean
  - Use the **transformation** $y_i = x_i$
  - Take the arithmetic mean of all y_i s to get $\bar{y}$
  - Finally, $\bar{x} = \bar{y}$

- For geometric mean
  - Use the **transformation** $\boldsymbol{y_i = \log(x_i)}$
  - Take the arithmetic mean of all y_i s to get $\bar{y}$
  - Finally, $\hat{x} = e^{\bar{y}}$

- For harmonic mean
  - Use the **transformation** $\boldsymbol{y_i = \frac{1}{x_i}}$
  - Take the arithmetic mean of all y_i s to get $\bar{y}$
  - Finally, $\tilde{x} = \frac{1}{\bar{y}}$

RELATIONSHIP AMONG MEANS

- A simple inequality exists between the three means related summary measure as

$$AM \geq GM \geq HM$$

MEDIAN OF A SAMPLE

Definition 3.12: **Median of a sample**

Median of a sample is the middle value when the data are arranged in increasing (*or decreasing*) order. Symbolically,

$$\hat{x} = \begin{cases} x_{(n+1)/2} & \textit{if } n \textit{ is odd} \\ \frac{1}{2}\left\{x_{n/2} + x_{(\frac{n}{2}+1)}\right\} & \textit{if } n \textit{ is even} \end{cases}$$

MEDIAN OF A GROUPED DATA

Definition 3.12: **Median of a grouped data**

Median of a grouped data is given by

$$\hat{x} = l + \left\{ \frac{\frac{N}{2} - cf}{f} \; h \right\}$$

where h = width of the median class

$N = \sum_{i=1}^{n} f_i$

f_i is the frequency of the i^{th} class, and n is the total number of groups

cf = the cumulative frequency (previous of the modal class)

f = the frequency of the particular class group

l = lower limit of the median class

Note

A class is called median class if its cumulative frequency is just greater than $N/2$

- Find the median of the grouped data representing yearly income of number of persons.

Income	No. of persons
60-69	5
70-79	15
80-89	20
90-99	30
100-109	20
110-119	8

Step1 : Exclusive group

Income	Exclusive	No. of persons	cf.
60-69	59.5 - 69.5	5	5
70-79	69.5 - 79.5	15	20
80-89	79.5 - 89.5	20	40
90-99	89.5 - 99.5	30	70
100-109	99.5 - 109.5	20	90
110-119	109.5 - 119.5	8	98

$$\text{Median} = l + \frac{\frac{N}{2} - c.f}{f} \times h.$$

$$= 89.5 + \left(\frac{49 - 40}{30}\right) 10$$

Median = 92.5

$$\frac{N}{2} = \frac{98}{2} = 49.$$

$l = 49.$

$c.f = 40.$

$f = 30.$

$h = 10.$

MODE OF A SAMPLE

- Mode is defined as the observation which occurs most frequently.
- For example, number of wickets obtained by bowler in 10 test matches are as follows.

1 2 0 3 2 4 1 1 2 2

- In other words, the above data can be represented as:-

	0	1	2	3	4
# of matches	1	3	4	1	1

- Clearly, the mode here is “2”.

MODE OF A GROUPED DATA

Definition 3.13: **Mode of a grouped data**

Select the modal class (it is the class with the highest frequency). Then the mode $\tilde{x}$ is given by:

$$\tilde{x} = l + \left(\frac{\Delta_1}{\Delta_1 - \Delta_2}\right) h$$

where,

h is the class width

Δ_1 is the difference between the frequency of the modal class and the frequency of the class preceding the modal class

Δ_2 is the difference between the frequency of the modal class and the class succeeding the modal class

l is the lower boundary of the modal class

Note

If each data value occurs only once, then there is no mode!

Mode of Grouped Data.

Marks	Frequency.	c.f.
20 - 30	4	
30 - 40	11	
40 - 50	18	
50 - 60	30	
60 - 70	27	
70 - 80	10.	

Modal class : Highest frequency. ⟶ 30

f_1 : frequency of modal class.

f_0 : frequency of previous class. ⟶ 18

f_2 : frequency of succeeding class. ⟶ 27.

l : lower limit of modal class. ⟶ 50.

h : class interval. ⟶ 10.

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \cdot h$$

$$= 50 + \left(\frac{30 - 18}{2(30) - 18 - 27}\right) \cdot 10.$$

Mode = 58

$$\tilde{x} = l + \left(\frac{\Delta_1}{\Delta_1 - \Delta_2}\right) h$$

$$= l + \left(\frac{f_1 - f_0}{(f_1 - f_0) - (f_2 - f_1)}\right) h$$

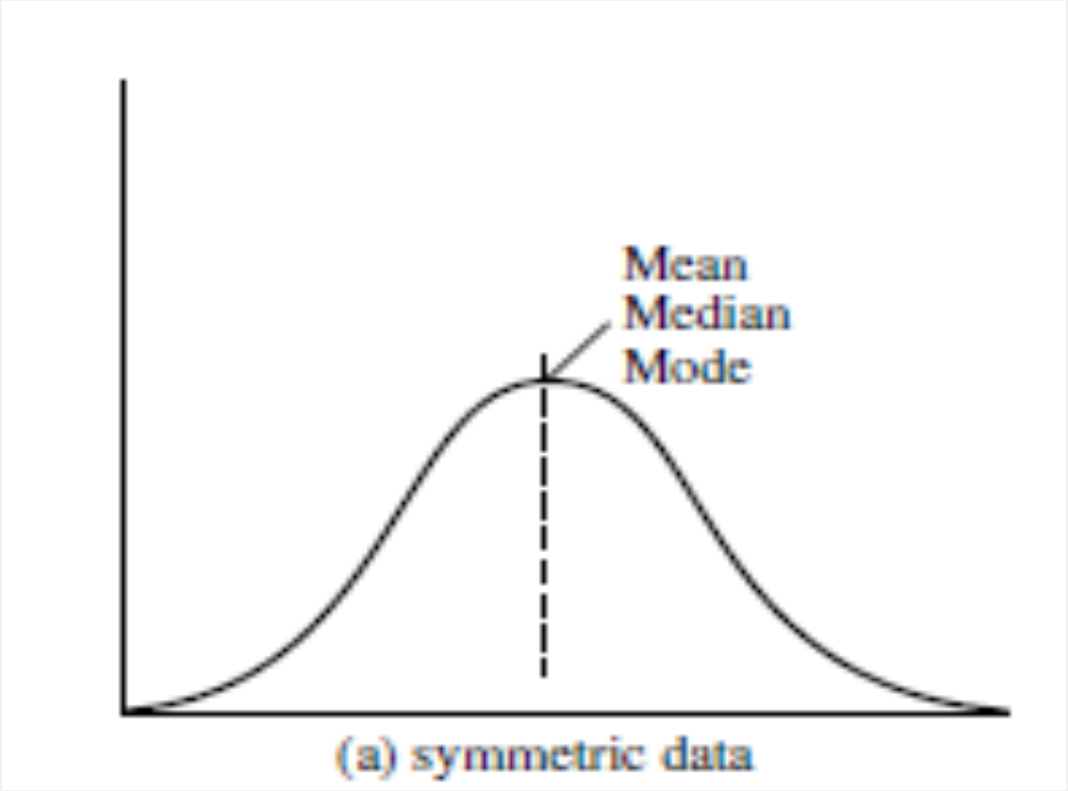
$$= = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)$$

RELATION BETWEEN MEAN, MEDIAN AND MODE

- A given set of data can be categorized into three categories:-
  - Symmetric data
  - Positively skewed data
  - Negatively skewed data

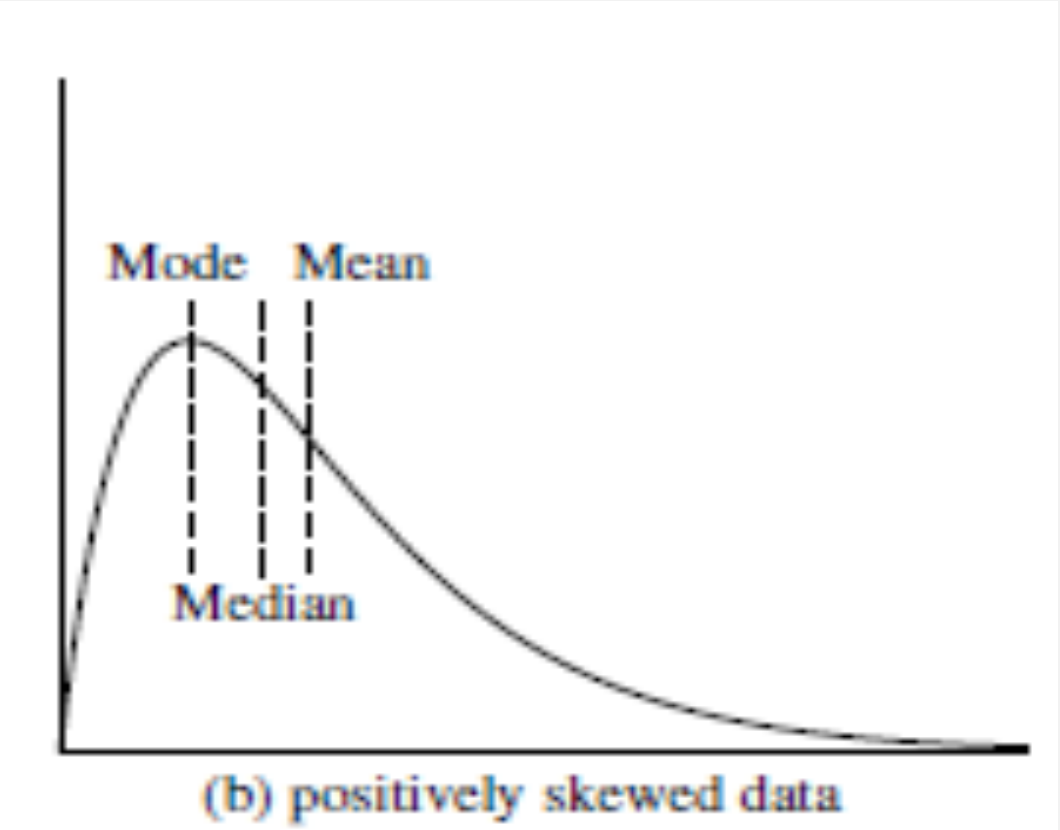
SYMMETRIC DATA

- For symmetric data, all mean, median and mode lie at the same point



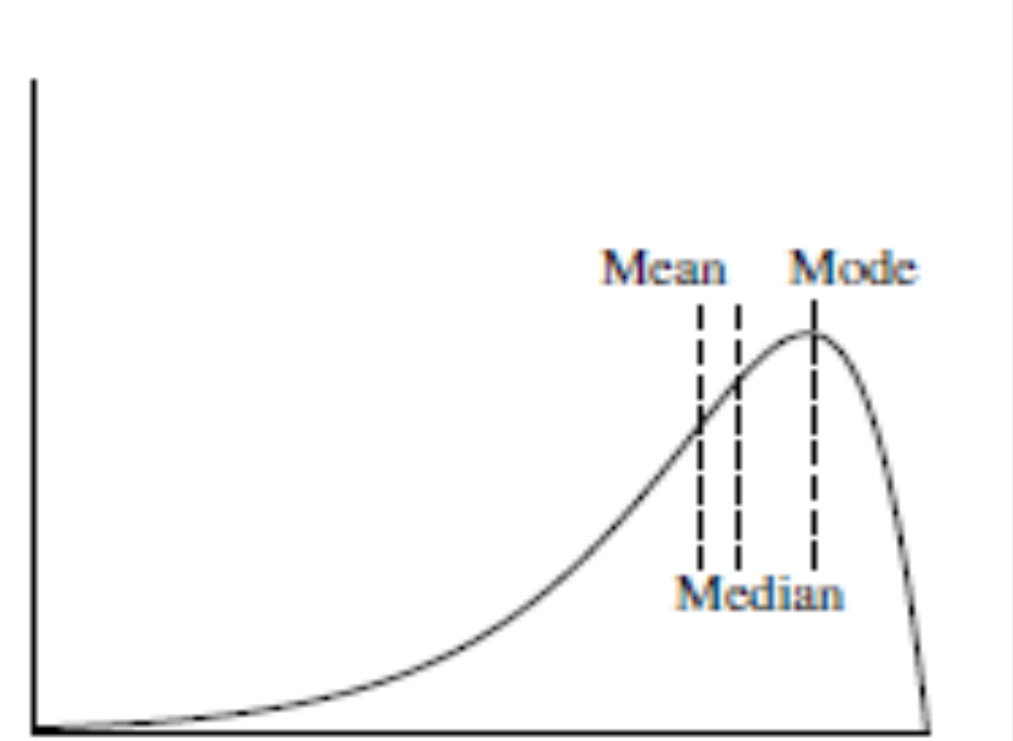
POSITIVELY SKEWED DATA

- Here, mode occurs at a value smaller than the median



NEGATIVELY SKEWED DATA

- Here, mode occurs at a value greater than the median



(c) negatively skewed data

EMPIRICAL RELATION!

- There is an empirical relation, valid for moderately skewed data

$$Mean - Mode = 3 * (Mean - Median)$$

MIDRANGE

- It is the average of the largest and smallest values in the set.

Find the midrange for the following set of numbers:

2, 4, 7, 10, 14, 35

Midrange = ?

EXAMPLES

1. **Raghav received the following scores on his mathematics exams: 84, 92, 74, 98, and 82. Find the mean, median, and mode of his scores.**

2. **During a seven-day period in July, a meteorologist recorded that the median daily high temperature was 91°. Which of the following are true statements?**

i) **The high temperature was exactly 91° on each of the seven days.**

ii) **The high temperature was never lower than 92°.**

iii) **Half the high temperatures were above 91° and half were below 91°.**

A) i only

B) ii only

C) iii only

D) i, ii, and iii

MEASURES OF DISPERSION

- Location measure are far too insufficient to understand data.
- Another set of commonly used summary statistics for continuous data are those that measure the dispersion.
- A dispersion measures the extent of spread of observations in a sample.

- Some important measure of dispersion are:
  - Range
  - Variance and Standard Deviation
  - Mean Absolute Deviation (MAD)
  - Absolute Average Deviation (AAD)
  - Interquartile Range (IQR)

MEASURES OF DISPERSION

Example

- Suppose, two samples of fruit juice bottles from two companies ***A*** and ***B***. The unit in each bottle is measured in litre.

Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

- Both samples have same mean. However, the bottles from company A with more uniform content than company B.
- We say that the dispersion (or variability) of the observation from the average is less for A than sample B.
  - The variability in a sample should display how the observation spread out from the average
  - In buying juice, customer should feel more confident to buy it from A than B

RANGE OF A SAMPLE

Definition 3.14: **Range of a sample**

Let $\mathbf{X = x_1, x_2, x_1, \ldots\ldots, x_n}$ be **n** sample values that are arranged in increasing order.

The range **R** of these samples are then defined as:

$$\mathbf{R = max(X) - min(X) = x_n - x_1}$$

- Range identifies the maximum spread, it can be misleading if most of the values are concentrated in a narrow band of values, but there are also a relatively small number of more extreme values.

- The variance is another measure of dispersion to deal with such a situation.

VARIANCE AND STANDARD DEVIATION

Definition 3.15: **Variance and Standard Deviation**

Let $\mathbf{X} = \{ \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_1}, \ldots\ldots, \mathbf{x_n} \}$ are sample values of **n** samples. Then, variance denoted as $\boldsymbol{\sigma^2}$ is defined as :-

$$\boldsymbol{\sigma^2} = \frac{\mathbf{1}}{\mathbf{n-1}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \bar{\mathbf{x}})^2$$

where, $\bar{x}$ denotes the mean of the sample

The standard deviation, $\boldsymbol{\sigma}$, of the samples is the square root of the variance $\boldsymbol{\sigma^2}$

COEFFICIENT VARIATION

- **Basic properties**
  - σ measures spread about mean and should be chosen only when the mean is chosen as the measure of central tendency
  - $\sigma = 0$ only when there is no spread, that is, when all observations have the same value, otherwise $\sigma > 0$

Definition 3.16: **Coefficient variation**

A related measure is the coefficient of variation **CV**, which is defined as follows

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

This gives a ratio measure to spread.

MEAN ABSOLUTE DEVIATION (MAD)

- Since, the mean can be distorted by outlier, and as the variance is computed using the mean, it is thus sensitive to outlier. To avoid the effect of outlier, there are two more robust measures of dispersion known. These are:

  - Mean Absolute Deviation (MAD)

$$\mathbf{MAD\ (X) = median\ (\{|x_1 - \bar{x}|, \ldots\ldots, |x_n - \bar{x}|\})}$$

  - Absolute Average Deviation (AAD)

$$\mathbf{AAD(X) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|}$$

where, X = $\{x_1, x_2, \ldots\ldots, x_n\}$ *is the sample values of n observations*

INTERQUARTILE RANGE

- Like MAD and AAD, there is another robust measure of dispersion known, called as Interquartile range, denoted as IQR
- To understand IQR, let us first define *percentile* and *quartile*
- **Percentile**
  - The percentile of a set of ordered data can be defined as follows:
    - Given an ordinal or continuous attribute **x** and a number **p** between 0 and 100, the $\mathbf{p^{th}}$ percentile $\mathbf{x_p}$ is a value of **x** such that **p**% of the observed values of **x** are less than $\mathbf{x_p}$
    - Example: The $\mathbf{50^{th}}$ percentile is that value $\mathbf{x_{50\%}}$ such that **50%** of all values of **x** are less than $\mathbf{x_{50\%}}$.
  - **Note:** The median is the $\mathbf{50^{th}}$ percentile.

INTERQUARTILE RANGE

- **Quartile**
  - The most commonly used percentiles are quartiles.
    - The first quartile, denoted by $\mathbf{Q_1}$ is the $\mathbf{25^{th}}$ percentile.
    - The third quartile, denoted by $\mathbf{Q_3}$ is the $\mathbf{75^{th}}$ percentile
    - The median, $\mathbf{Q_2}$ is the $\mathbf{50^{th}}$ percentile.

- The quartiles including median, give some indication of the center, spread and shape of a distribution.

- The distance between $\mathbf{Q_1}$ and $\mathbf{Q_3}$ is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (**IQR**) and is defined as

$$\mathbf{IQR = Q_3 - Q_1}$$

APPLICATION OF IQR

- **Outlier detection using five-number summary**

  - A common rule of the thumb for identifying suspected outliers is to single out values falling at least **1.5 × IQR** above $\mathbf{Q_3}$ and below $\mathbf{Q_1}$.

  - In other words, extreme observations occurring within **1.5 × IQR** of the quartiles

APPLICATION OF IQR

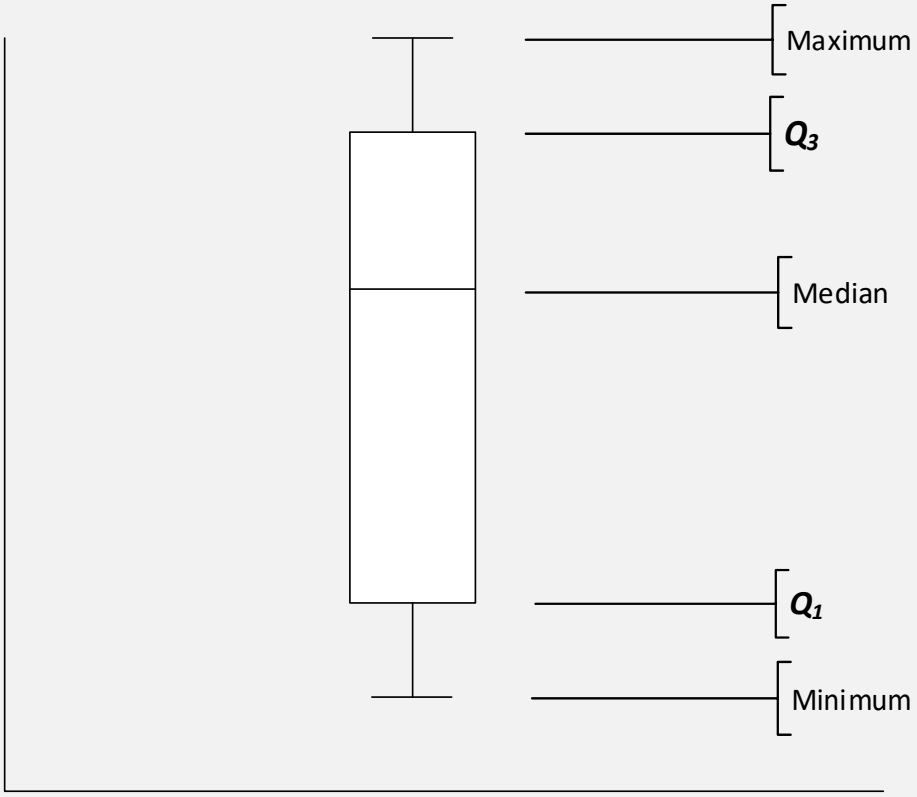
- **Five Number Summary**
  - Since, $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ together contain no information about the endpoints of the data, a complete summary of the shape of a distribution can be obtained by providing the lowest and highest data value as well. This is known as the five-number summary
  - The five-number summary of a distribution consists of :
    - The Median $\mathbf{Q_2}$
    - The first quartile $\mathbf{Q_1}$
    - The third quartile $\mathbf{Q_3}$
    - The smallest observation
    - The largest observation

These are, when written in order gives the **five-number summary**:

Minimum, $\mathbf{Q_1}$, Median ($\mathbf{Q_2}$), $\mathbf{Q_3}$, Maximum

BOX PLOT

- **Graphical view of Five number summary**



REFERENCE

- The detail material related to this lecture can be found in

Probability and Statistics for Enginneers and Scientists (8th Ed.) by Ronald E. Walpol, Sharon L. Myers, Keying Ye (Pearson), 2013 .

QUESTIONS OF THE DAY…

1. Which of the following central tendency measurements allows distributive, algebraic and holistic measure?
   - mean
   - median
   - Mode

   Which measure may be faster than other? Why?
2. Give three situations where AM, GM and HM are the right measure of central tendency?

QUESTIONS OF THE DAY…

3. Given a sample of data, how to decide whether it is
   a) Symmetric?
   b) Skew-symmetric (positive or negative)?
   c) Uniformly increasing (or decreasing)?
   d) In-variate?

4. How the box-plots will look for the following types of samples?
   a) Symmetric
   b) Positively skew-symmetric
   c) Negatively skew-symmetric
   d) in-variate

QUESTIONS OF THE DAY…

5. Draw the curves for the following types of distributions and clearly mark the likely locations of mean, median and mode in each of them.
   a. Symmetric
   b. Positively skew-symmetric
   c. Negatively skew-symmetric

6. The variance $\boldsymbol{\sigma^2}$ of a sample $\mathbf{X} = \{ \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_1}, \ldots\ldots, \mathbf{x_n} \}$ of $\boldsymbol{n}$ data is defined as follows.

$$\boldsymbol{\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where, $\bar{x}$ denotes the mean of the sample. Why *(n-1)* is in the denominator in stead of *n*?

QUESTIONS OF THE DAY…

7. What are the degree of freedoms in each of the following cases.
    a. A sample with a single data
    b. A sample with *n* data
    c. A sample of tabular data with *n* rows and *m* columns