Data Analytics (CS61061)

Lecture #3

Descriptive Statistics

Dr. Debasis Samanta

Professor

Department of Computer Science & Engineering

Quote of the day..

- Change your thoughts and you change your world.
  - NORMAN VINCENT PEALE, American - Clergyman

Today's discussion includes…

- Introduction
- Data summarization
  - 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

TRP: An example

- Television rating point (TRP) is a tool provided to judge which programs are viewed the most.
  - This gives us an index of the choice of the people and also the popularity of a particular channel.
- For calculation purpose, a device is attached to the TV sets in few thousand viewers' houses in different geographic and demographic sectors.
  - The device is called as **People's Meter**. It reads the time and the program that a viewer watches on a particular day for a certain period.
- An average is taken, for example, for a 30-days period.
- The above further can be augmented with a personal interview survey (PIS), which becomes the basis for many studies/decision making.
- Essentially, we are to analyze **data** for TRP estimation.

Defining Data

Definition 3.1: **Data**

A set of data is a collection of observed values representing one or more characteristics of some objects or units.

Example: For TRP, data collection consist of the following attributes.

- **Age:** A viewer's age in years
- **Sex:** A viewer's gender coded 1 for male and 0 for female
- **Happy:** A viewer's general happiness
  - NH for not too happy
  - PH for pretty happy
  - VH for very happy
- **TVHours:** The average number of hours a respondent watched TV during a day

Defining Data

Viewer#	Age	Sex	Happy	TVHours
…	…	…	…	…
…	…	…	…	…
55	34	F	VH	5
…	…	…	…	…

Note:

- A data set is composed of information from a set of units.
- Information from a unit is known as an observation.
- An observation consists of one or more pieces of information about a unit; these are called variables.

Defining Population

Definition 3.2: **Population**

A population is a data set representing the entire entities of interest.

Example: All TV viewers in the country/world.

Note:

1. All people in the country/world is not a population.
2. For different survey, the population set may be completely different.
3. For statistical learning, it is important to define the population that we intend to study very carefully.

Defining Sample

Definition 3.3: **Sample**

A sample is a data set consisting of a population.

Example: All students studying in Class XII is a sample, whereas those students belong to a given school is population.

Note:

- Normally a sample is obtained in such a way as to be representative of the population.

Defining Statistics

Definition 3.4: **Statistics**

A statistics is a quantity calculated from data that describes a particular characteristics of a sample.

Example: The sample mean (denoted by $\bar{y}$) is the arithmetic mean of a variable of all the observations of a sample.

Defining Statistical Inference

Definition 3.5: **Statistical inference**

Statistical inference is the process of using sample statistics to make decisions about population.

Example: In the context of TRP

- Overall frequency of the various levels of happiness.
- Is there a relationship between the age of a viewers and his/her general happiness?
- Is there a relationship between the age of the viewer and the number of TV hours watched?

Data Summarization

- To identify the typical characteristics of data (i.e., to have an overall picture).
- To identify which data should be treated as noise or outliers.
- The data summarization techniques can be classified into two broad categories:
  - Measures of **location**
  - Measures of **dispersion**

Measurement of location

- It is also alternatively called as measuring the central tendency.
  - A function of the sample values that summarizes the location information into a single number is known as a measure of location.
- The most popular measures of location are
  - Mean
  - Median
  - Mode
  - Midrange
- These can be measured in three ways
  - Distributive measure
  - Algebraic measure
  - Holistic measure

Distributive measure

- It is a measure (*i.e.*, *function*) that can be computed for a given data set by partitioning the data into smaller subsets, computing the measure for each subset, and then merging the results in order to arrive at the measure's value for the original (*i.e.*, *entire*) data set.

Example

- ✓ sum(), count()

Algebraic measure

- It is a measure that can be computed by applying an algebraic function to one or more distributive measures.

- **Example**

$$\text{average} = \frac{\text{sum}()}{\text{count}()}$$

Holistic measure

- It is a measure that must be computed on the entire data set as a whole.

- **Example**

  Calculating median

  What about *mode?*

Arithmetic Mean

Mean of a sample

- The mean of a sample data is denoted as $\bar{x}$. Different mean measurements known are:
  - Simple mean
  - Weighted mean
  - Trimmed mean
- In the next few slides, we shall learn how to calculate the mean of a sample.
- We assume that given $x_1, x_2, x_3, \ldots, x_n$ are the sample values.

Simple mean of a sample

- **Simple mean**

  It is also called simply arithmetic mean or average and is abbreviated as (AM) and denoted as $\bar{x}$.

Definition 3.6: **Simple mean**

- ✓ If $x_1, x_2, x_3, \ldots, x_n$ are the sample values, the simple mean is defined as

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

Weighted mean of a sample

- **Weighted mean**

  It is also called weighted arithmetic mean or weighted average.

Definition 3.7: **Weighted mean**

When each sample value x_i is associated with a weight w_i, for $i = 1,2,\ldots,n$, then it is defined as

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

Note

When all weights are equal, the weighted mean reduces to simple mean.

Trimmed mean of a sample

- **Trimmed Mean**

  If there are extreme values (*also called outlier*) in a sample, then the mean is influenced greatly by those values. To offset the effect caused by those extreme values, we can use the concept of trimmed mean

Definition 3.8: **Trimmed mean**

Trimmed mean is defined as the mean obtained after chopping off values at the high and low extremes.

Properties of mean

- **Lemma 3.1**

  If $\bar{x}_i$, $i = 1,2,\ldots,m$ are the means of m samples of sizes $n_1, n_2, \ldots, n_m$ respectively, then the mean of the combined sample is given by

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i}$$

  (Distributive Measure)

- **Lemma 3.2**

  - ✓ If a new observation x_k is added to a sample of size n with mean $\bar{x}$, the new mean is given by

$$\bar{x}' = \frac{n\bar{x} + x_k}{n+1}$$

Properties of mean

- **Lemma 3.3**

  If an existing observation x_k is removed from a sample of size n with mean $\bar{x}$, the new mean is given by

$$\bar{x}' = \frac{n\,\bar{x} - x_k}{n-1}$$

- **Lemma 3.4**

  If m observations with mean $\bar{x}_m$, are added (*removed*) from a sample of size n with mean $\bar{x}_n$, then the new mean is given by

$$\bar{x} = \frac{n\,\bar{x}_n \pm m\,\bar{x}_m}{n \pm m}$$

Properties of mean

- **Lemma 3.5**

  If a constant ***c*** is subtracted (*or added*) from each sample value, then the mean of the transformed variable is linearly displaced by *c*. That is,

$$\bar{x}' = \bar{x} \mp c$$

- **Lemma 3.6**

  If each observation is called by multiplying (*dividing*) by a non-zero constant, then the altered mean is given by

$$\bar{x}' = \bar{x} * c$$

  *Where, * is* ***x*** *(multiplication) or ÷ (division) operator.*

Mean with grouped data

Sometimes data is given in the form of classes and frequency for each class.

Class →	x_1- x_2	x_2- x_3	.....	x_i- x_{i+1}	.....	x_{n-1}- x_n
Frequency →	f_1	f_2	.....	f_i	.....	f_n

There three methods to calculate the mean of such a grouped data.

- Direct method
- Assumed mean method
- Step deviation method

Direct method

- **Direct Method**

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

Where, $\mathbf{x_i} = \frac{1}{2}$ **(lower limit + upper limit)** of the i^{th} class, i.e., $x_i = \frac{x_i + x_{i+1}}{2}$ (also called class size), and f_i is the frequency of the i^{th} class.

Note

$$\sum f_i (x_i - \bar{x}) = 0$$

Assumed mean method

- **Assumed mean method**

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}$$

where, A is the assumed mean (it is usually a value $x_i = \frac{x_i + x_{i+1}}{2}$ chosen in the middle of the groups, and $d_i = (A - x_i)$ for each i .

Step deviation method

- **Step deviation method**

$$\bar{x} = A + \left\{ \frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} \; h \right\}$$

where,

A = assumed mean

h = class size (*i.e.*, $\mathbf{x_{i+1} - x_i}$ *for the* i^{th} *class*)

$u_i = \frac{x_i - A}{h}$

Mean for a group of data

- For the above methods, we assume that…
  - All classes are equal sized
  - Groups are with inclusive classes, i.e., $\mathbf{x_i = x_{i-1}}$(*linear limit of a class is same as the upper limit of the previous class*)

10 - 19	20 - 29	30 - 39	40 – 49

*Data with **ex**clusive classes*

9.5 – 19.5	19.5 – 29.5	29.5 – 39.5	39.5 – 49.5

*Data with **in**clusive classes*

Ogive: Graphical method to find mean

- **Ogive** (pronounced as O-Jive) is a cumulative frequency polygon graph.
  - When cumulative frequencies are plotted against the upper (lower) class limit, the plot resembles one side of an Arabesque or **ogival** architecture, hence the name.
  - There are two types of Ogive plots
    - Less-than (upper class versus cumulative frequency)
    - More than (lower class versus cumulative frequency)

Example:

Suppose, there is a data relating the marks obtained by 200 students in an examination

444, 412, 478, 467, 432, 450, 410, 465, 435, 454, 479, ...

(Further, suppose it is observed that the minimum and maximum marks are 410, 479, respectively.)

Ogive: Cumulative frequency table

444, 412, 478, 467, 432, 450, 410, 465, 435, 454, 479, ...

Step 1: Draw a cumulative frequency table

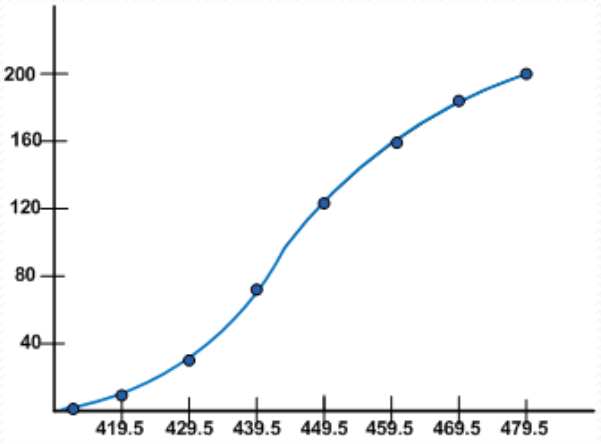
Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

Ogive: Graphical method to find mean

Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

Step 2: Less-than Ogive graph

Upper class	Cumulative Frequency
Less than 419.5	14
Less than 429.5	34
Less than 439.5	76
Less than 449.5	130
Less than 459.5	175
Less than 469.5	193
Less than 479.5	200

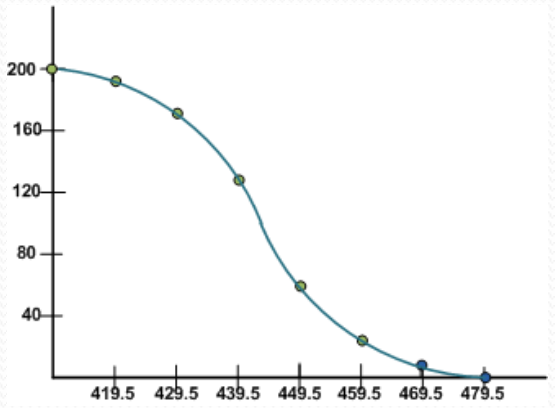


Ogive: Graphical method to find mean

Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

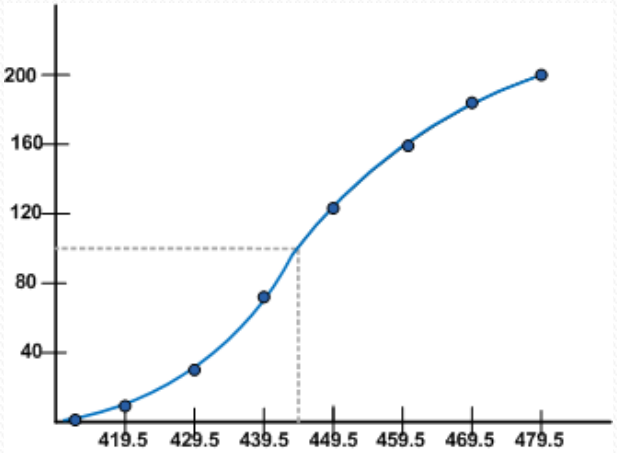
Step 3: More-than Ogive graph

Lower class	Cumulative Frequency
More than 409.5	200
More than 419.5	186
More than 429.5	166
More than 439.5	124
More than 449.5	70
More than 459.5	25
More than 469.5	7

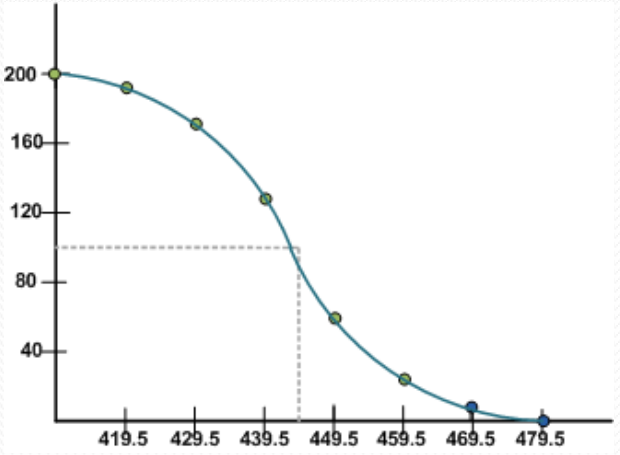


Information from Ogive

- Mean from Less-than Ogive



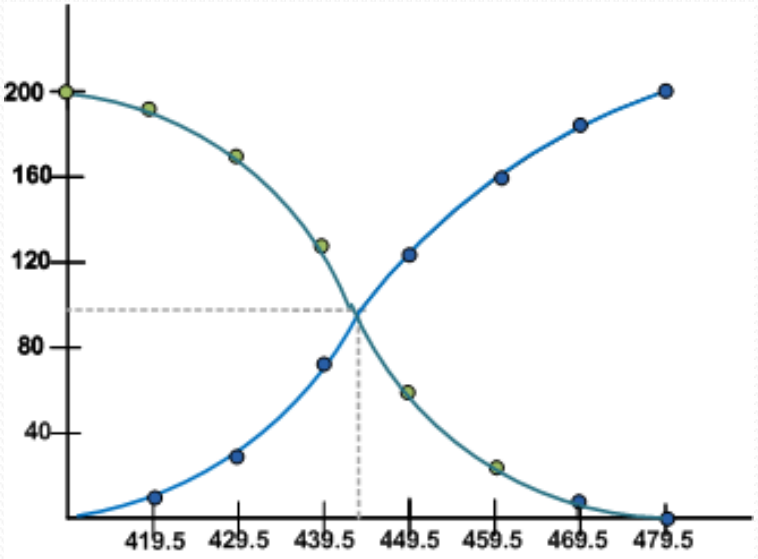
- Mean from More-than Ogive



- A % C freq of .65 for the third class 439.5.....449.5 means that 65% of all scores are found in this class or below.

Information from Ogive

- Less-than and more-than Ogive approach



A cross point of two Ogive plots gives the mean of the sample

Some other measures of mean

- There are three mean measures of location:
  - Arithmetic Mean (AM)
  - Geometric mean (GM)
  - Harmonic mean (HM)

These three means are called Pythagorean means

Some other measures of mean

- Arithmetic Mean (**AM**)
  - $S: \{x_1, x_2\}$
  - $\bar{x} = \frac{x_1 + x_2}{2}$
  - $\bar{x} - x_1 = x_2 - \bar{x}$
- Geometric mean (**GM**)
  - $S: \{x_1, x_2\}$
  - $\tilde{x} = \sqrt{x_1 . x_2}$
  - $\frac{x_1}{\tilde{x}} = \frac{\tilde{x}}{x_2}$
- Harmonic Mean (**HM**)
  - $S: \{x_1, x_2\}$
  - $\hat{x} = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$
  - $\frac{2}{\hat{x}} = \frac{1}{x_1} + \frac{1}{x_2}$

???

- 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.

Usefulness of different means

Applications of GM

- The geometric mean is most useful to calculate the compounded growth rate where values in the sample are **not** independent of each other

Example: Calculation of growth rate

Consider a stock that grows by 10% in year one, declines by 20% in year two, and then grows by 30% in year three. What is the growth rate?

$$GM = \sqrt[3]{(1+0.1)(1-0.2)(1+0.3)}$$

$$= 0.046$$

= 4.6% annually.

Applications of GM

- Other some applications:
  - if values tend to make large fluctuations.
  - when the values that are multiplied together are exponential value
    - the statistical rates of human population growth in consecutive 10 or 20 years, etc.

Why Geometric Mean is “geometric”?

Applications of HM

- In fact, harmonic mean is the reciprocal of mean of sum of reciprocals
  - Harmonic means are often used in averaging things like rates.

Example: Calculate travel speed given durations of several trips

- A car in first 1 hour travels 60 kmph, in next 2 hours it travels with 90 kmph and next in three hours it travels with 80 kmph

Calculation using AM:

$$\text{Average speed} = \frac{Total\ disatnec}{Total\ time} = \frac{60+180+240}{1+2+3} = \frac{480}{6} = 80\ kmph$$

Calculation using HM

$$\text{Average speed} = \frac{1+2+3}{\frac{1}{60}+\frac{2}{90}+\frac{3}{80}} = \frac{6}{0.0167+0.0222+0.0375} = \frac{6}{0.0764} = 78.5\ kmph$$

Applications of weighted HM

- The weighted harmonic mean of x_1, x_2, x_3 with the corresponding weights w_1, w_2, w_3 is given as:

$$whm = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

Example: Calculation of price earning ratio (per)

Consider two firms: one has a market capitalization of $100 billion and earnings of $4 billion (per of 25) and another with a market capitalization of $1 billion and earnings of $4 million (per of 250). In an index made of the two stocks, with 10% invested in the first and 90% invested in the second, what is the per of the index?

Applications of weighted HM

- **Example: Calculation of price earning ratio (per)**

  Consider two firms: one has a market capitalization of $100 billion and earnings of $4 billion (per of 25) and another with a market capitalization of $1 billion and earnings of $4 million (per of 250). In an index made of the two stocks, with 10% invested in the first and 90% invested in the second, what is the per of the index?

  Calculation 1: Using weighted average mean

$$wam = 0.1 \times 25 + 0.9 \times 250 = 227.5$$

  Calculation 2: Using weighted harmonic mean

$$whm = \frac{0.1 + 0.9}{\frac{0.1}{25} + \frac{0.9}{250}} = 131.6$$

  As can be seen, the weighted arithmetic mean significantly overestimates the mean price-earnings ratio.

Other way of using means

Different mean calculations

- Let us consider a sample which considers two things
  - 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

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	4	…	7

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

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

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	4	…	7

$$\tilde{x} = \frac{n}{\sum_1^n \frac{1}{x_i}}$$

Here, $x_i = \frac{r_i}{d_i}$

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

$$\hat{x} = \left(\prod_{1}^{n} x_i \right)^{\frac{1}{n}} \qquad \text{where } x_i = r_i \times d_i$$

Key takeaways

- 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

Key takeaways

- 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** $y_i = \log(x_i)$
  - 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** $y_i = \frac{1}{x_i}$
  - Take the arithmetic mean of all y_i s to get $\bar{y}$
  - Finally, $\tilde{x} = \frac{1}{\bar{y}}$

Key takeaways

- All mean calculations are applicable to a set of data in a sample.
- In general, you can observe that

$$AM \geq GM \geq HM$$

- If your sample contains all values which are same (i.e., invariant), then

$$AM = GM = HM$$

- Apply AM, to find summary mean when data are in a normalized, GM to find growth rate and HM to find the summary mean when data are not normalized.

Median calculation

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,

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

Calculating median from a set of samples

- Consider the case of three sets of data:
  - Set 1: $x_{11}, x_{12}, x_{13}, \ldots, x_{1m}$
  - Set 2: $x_{21}, x_{22}, x_{23}, \ldots, x_{2n}$
  - Set 3: $x_{31}, x_{32}, x_{33}, \ldots, x_{3p}$

What is median?
Calculate in a memory-constrained computing environment.
Distributive/ algebraic/ holistic approach?

Median of a sample

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

N = the total number of samples

l = lower limit of the median class

Note

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

Mode calculation

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 width of the modal class

Δ_1 is the difference between the frequency of the modal class and the frequency of the class **just after** the modal class

Δ_2 is the difference between the frequency of the modal class and the class **just before** 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!

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

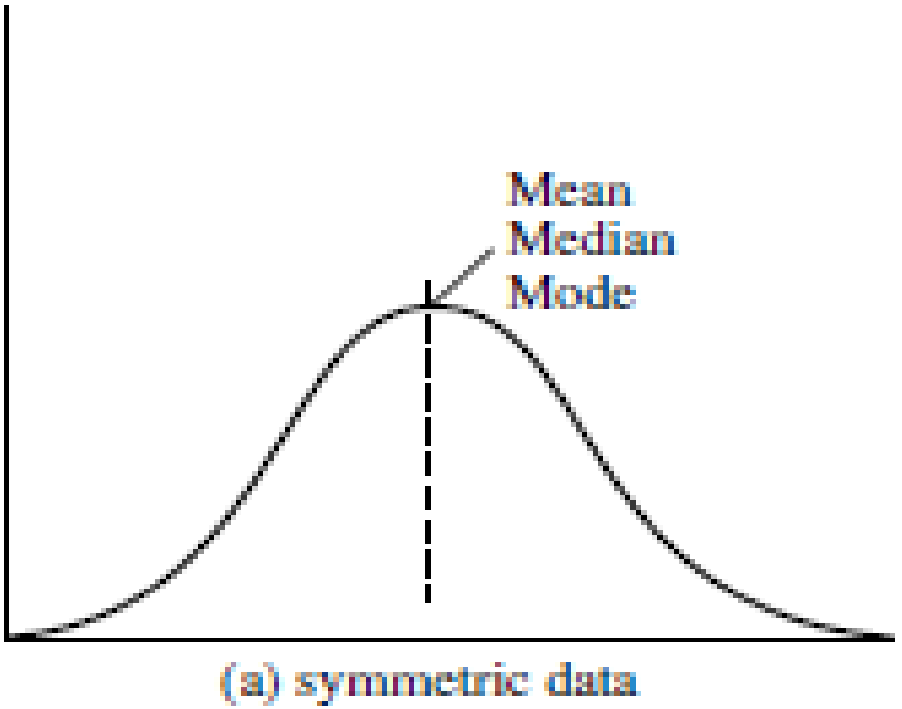
- To understand the above three categories, let us consider the following
- Given a set of $\boldsymbol{m}$ objects, where any object can take values $\boldsymbol{v_1, v_2, \ldots, v_k}$. Then, the frequency of a value $\boldsymbol{v_i}$ is defined as

$$\boldsymbol{Frequency(v_i) = \frac{Number\ of\ objects\ with\ value\ v_i}{n}}$$

$$for\ i = 1,2,\ldots,k$$

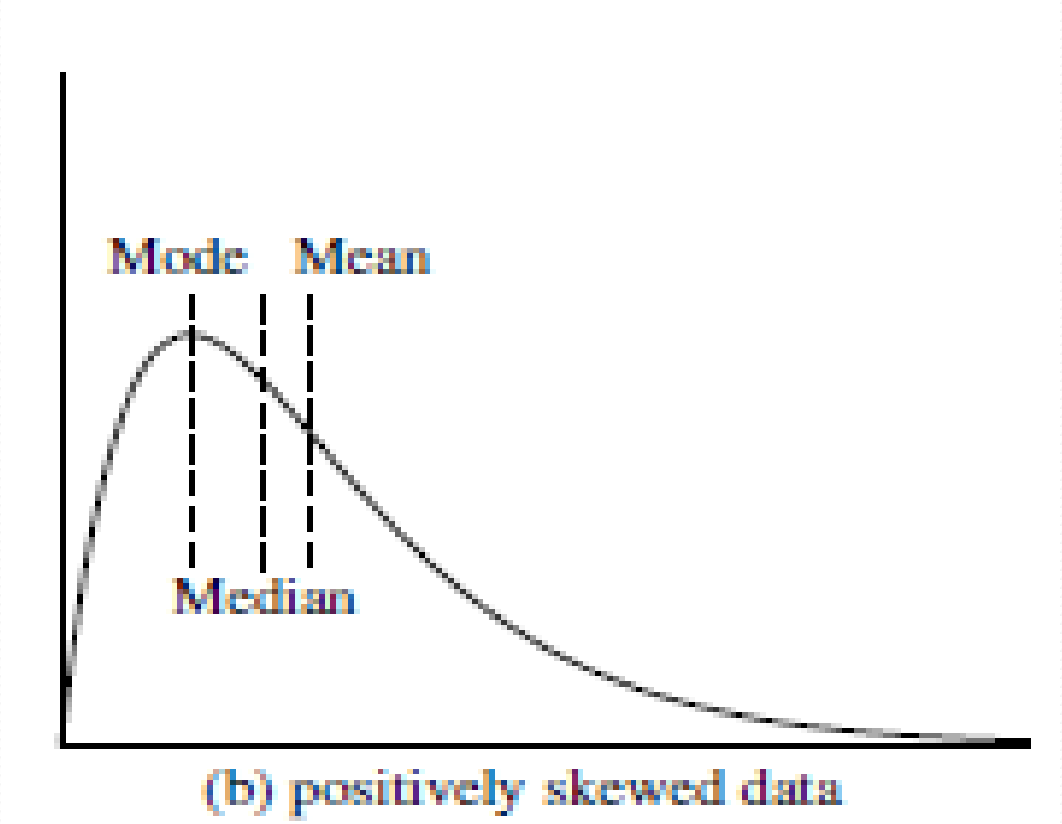
Symmetric data

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



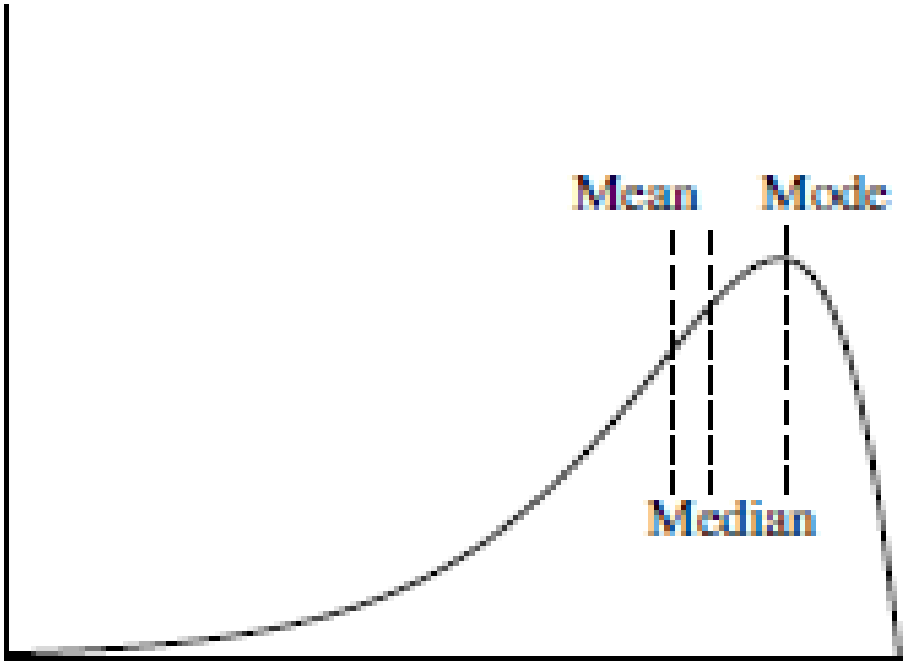
Positively skewed data

- 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!

- How to measure the skewness of data?

Skewness = 3 * (mean–median) / standard deviation

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

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

Midrange

- It is the trimmed mean of a range of values at the middle in a sample data set.

Steps for calculating midrange

1. A percentage ‘p’ between 0 and 100 is specified.
2. The top and bottom of (p/2)% of the data is thrown out.
3. The mean is then calculated in the normal way.

- Thus, the median is trimmed mean with p = 100% while the arithmetic mean corresponds to p = 0%

Note

- Trimmed mean is a special case of Midrange

Measure of dispersion

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 measures 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 $\mathbf{X = \{x_1, x_2, ..., x_n\}}$ denotes a sample of ***n*** 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 $\mathbf{X = \{ x_1, x_2, \ldots, x_n \}}$ is a sample of **n** data. Then, variance denoted as $\boldsymbol{\sigma^2}$ is defined as :-

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

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

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
  - **σ = 0** only when there is no spread, that is, when all observations have the same value, otherwise **σ > 0**

Definition 3.16: **Coefficient of variation**

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

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

This gives a ratio measure to spread.

Coefficient variation

- **Significance of CV**
  - It is a statistical measure of the dispersion of data points in a data series around the mean
  - **CV = p%** implies that standard deviation is p% to that of the mean of a sample.

Example:

- Suppose, there are three series of data representing amount of returns that investors receives from three farms F1, F2 and F3 in a year.
- CV(F1), CV(F2), CV(F3) indicate volatilities/ risks in comparison of return expected from investment

$$\mathbf{CV} = \frac{\mathbf{Volatilty}}{\mathbf{Expected\ Return}} \times \mathbf{100}$$

Variance and Standard Deviation

- **Lemma 3.8**
  - If data are transformed as $\mathbf{x' = \frac{(x-a)}{c}}$, the variance is transformed as

$$\sigma'^2 = \frac{1}{c^2}\sigma^2$$

Proof

The new mean $\bar{x}' = \frac{\bar{x}-a}{c}$

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i' - \bar{x}'\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(x_i-a)}{c} - \frac{(\bar{x}-a)}{c}\right\}^2$$

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

$$= \frac{1}{c^2 n}\sum_{i=1}^{n}\{x_i - \bar{x}\}^2$$

$$= \frac{1}{c^2}\sigma^2 \quad \text{[PROVED]}$$

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)

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

  - Absolute Average Deviation (AAD)

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

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 IQR measure**
  - 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

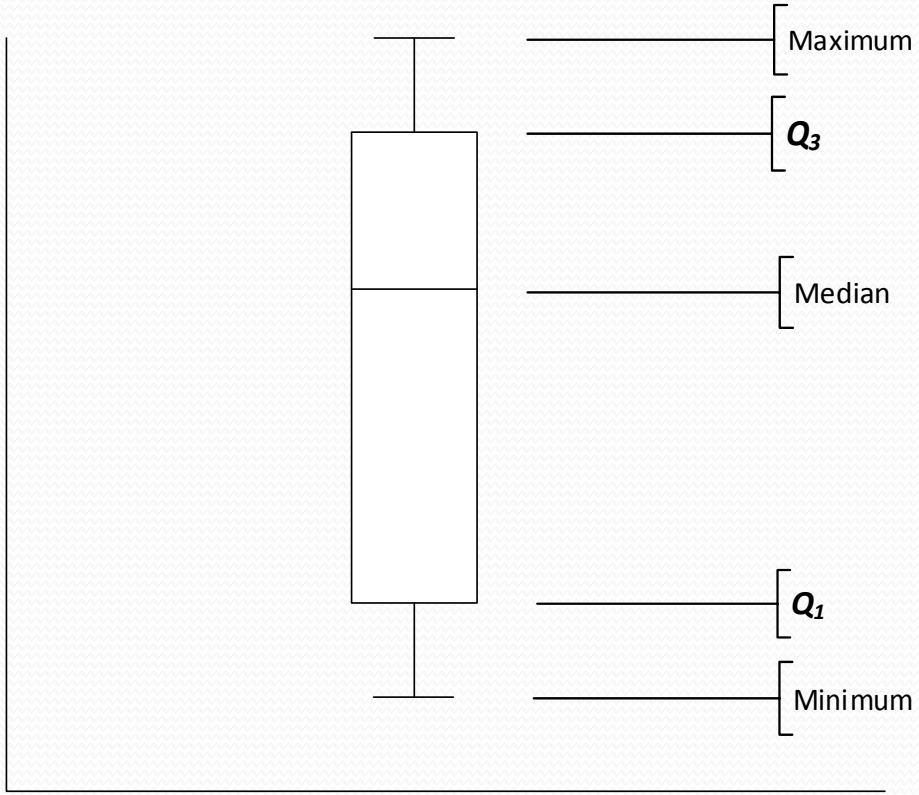
- **Five Number Summary**
  - Since, $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ together contain no information about the endpoints of the data, a complete summary of the shape of a distribution can be obtained by providing the lowest and highest data value as well. This is known as the five-number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

Any question?

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 measures 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 σ^2 of a sample $X = \{ x_1, x_2, x_1, \ldots, x_n \}$ of n data is defined as follows.

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

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

Questions of the day…

5. 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.

6. Calculate the Coefficient of variation (CV) for the following sample data.

   (a) 10, -5, 20, 15, -5, 25, 30, 35, -25, 25

   (b)

x	10	20	30	40	50
f(x)	0.2	0.4	0.1	0.2	0.1