Data Science (CS3206)

Lecture #4

Descriptive Statistics

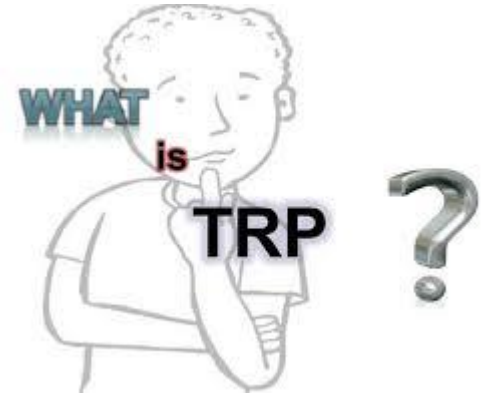
Quote of the day..

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

Today's discussion…

- 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 programme 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 : **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
- **Gender:** 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 : **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 : **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 : **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 : **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?*

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).

Definition : **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 : **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 : **Trimmed mean**

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

Properties of mean

- **Lemma 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 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**

  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 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 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 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* $\times$ *(multiplication) or* $\div$ *(division) operator.*

Mean with grouped data

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

Class			.....		.....	
Frequency			.....		.....	

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 $d_i = (A - x_i)$ for each i)

Assumed mean method

The following table gives the information about the marks obtained by 100 students in an examination.

Class	0-10	10-20	20-30	30-40	40-50
Frequency	12	28	32	25	13

Find the mean marks of the students using assumed mean method.

Solution:

Class (CI)	Frequency (f_i)	Class mark (x_i)	$d_i = x_i - a$	$f_i d_i$
0-10	12	5	5 – 25 = – 20	-240
10-20	28	15	15 – 25 = – 10	-280
20-30	32	25 = a	25-25 = 0	0
30-40	25	35	35-25 = 10	250
40-50	13	45	45-25 = 20	260
Total	Σf_i =100			$\Sigma f_i d_i$ = -10

Assumed mean = a = 25

Assumed mean method

Mean of the data:

$$\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i}$$

= 25 + (-10/ 100)

= 25 – 1/10

= (250-1)/10

= 249/10

=24.9

Hence, the mean marks of the students are 24.9.

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*)

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

Example

Mean for a group of data

- For the above methods, we can 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	

*Data with **ex**clusive classes*

9.5 – 19.5	19.5 – 29.5	29.5 – 39.5	

*Data with **in**clusive classes*

Some other measures of mean

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

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 : **Geometric mean**

Geometric mean of 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 : **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** $\boldsymbol{y_i = x_i}$
  - 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, $\hat{\boldsymbol{x}} = \boldsymbol{e^{\bar{y}}}$
- 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, $\tilde{\boldsymbol{x}} = \boldsymbol{\frac{1}{\bar{y}}}$

Relationship among means

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

$$AM \geq GM \geq HM$$

Median

Definition : **Median of a sample**

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

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

Median

Definition : **Median of a grouped data**

- Expensive to compute for a large number of observations.
- Let the interval that contains the median frequency be the median interval.
- Approximate the median of the entire data set by interpolation using the formula

$$median = L_1 + (\frac{N/2 - (\sum freq)_l}{freq_{median}})width$$

Where L_1 is lower boundary of median interval,

N is number of values in the entire data set

$(\sum freq)_l$ is the sum of frequencies of all the intervals that are lower than the median interval,

$freq_{median}$ is the frequency of the median interval,

width is the width of the median interval

Median

Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows.

age	*frequency*
1-5	200
5-15	450
15-20	300
20-50	1500
50-80	700
80-110	44

Compute an *approximate median* value for the data.

First identify median interval, L_1, N, width, $freq_{median}$, $(\sum freq)_l$

$$L_1 = 20,\ N = 3194,\ (\sum freq)_l = 950,\ freq_{median} = 1500,\ width = 30$$

Median = 32.94 years.

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 : **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 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

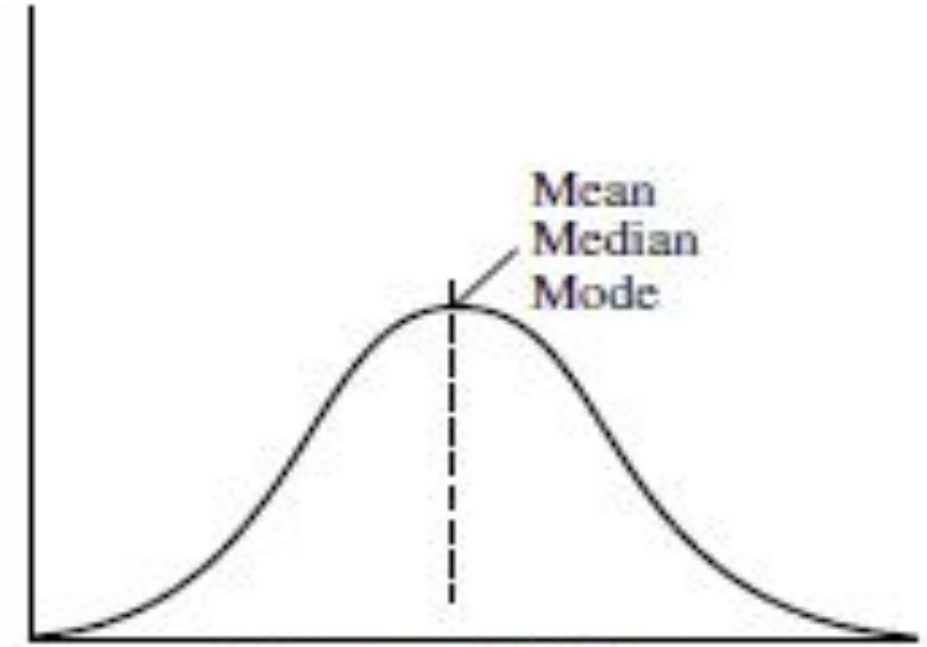
If each data value occurs only once, then there is no mode!

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

Symmetric data

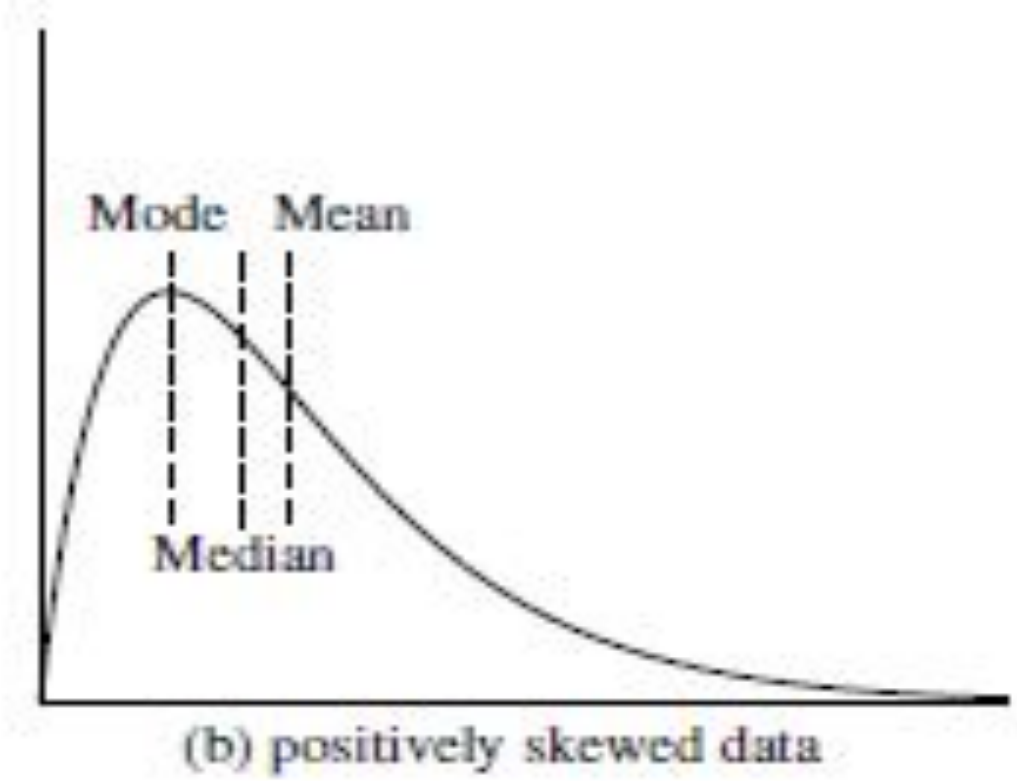
- For symmetric data, all mean, median and mode lie at the same point



(a) symmetric data

Positively skewed data

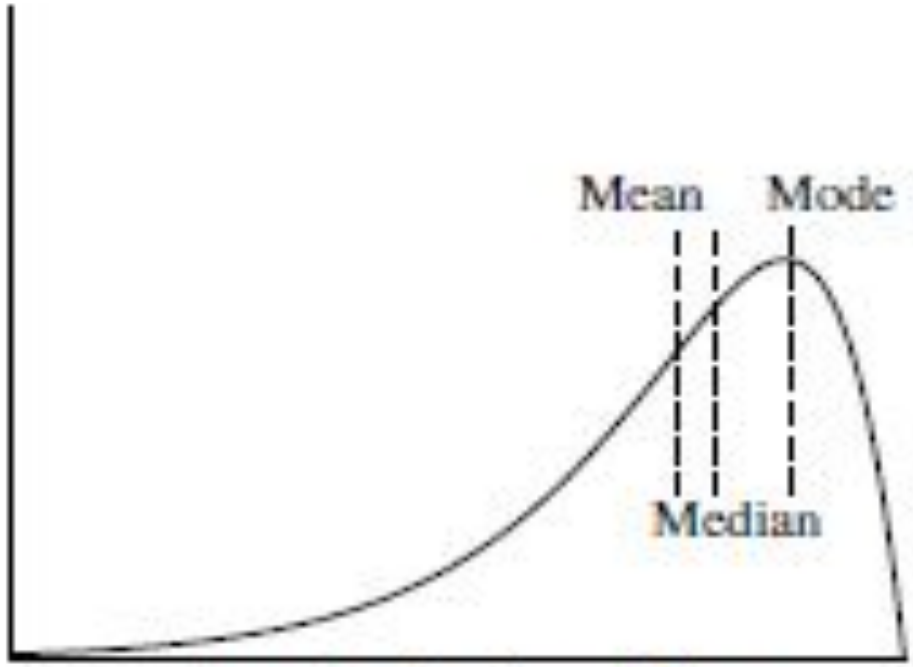
- Here, mode occurs at a value smaller than the median



(b) positively skewed data

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 unimodal moderately skewed data

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

The mode for unimodal frequency curves that are moderately skewed can easily be approximated if the mean and median values are known

Midrange

- The midrange can also be used to assess the central tendency of a numeric data set

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

Example

Suppose that the data for analysis includes the attribute *age. The age values for the data* tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the *mean of the data? What is the median?*

(b) What is the *mode of the data? Comment on the data's modality (i.e.bimodal,* trimodal, etc.).

(c) What is the *midrange of the data?*

a) Mean=30 , median=25
b) Bimodal 25 and 35
c) Midrange= 41.5

Data Science (CS3206)

Lecture #5

Descriptive Statistics

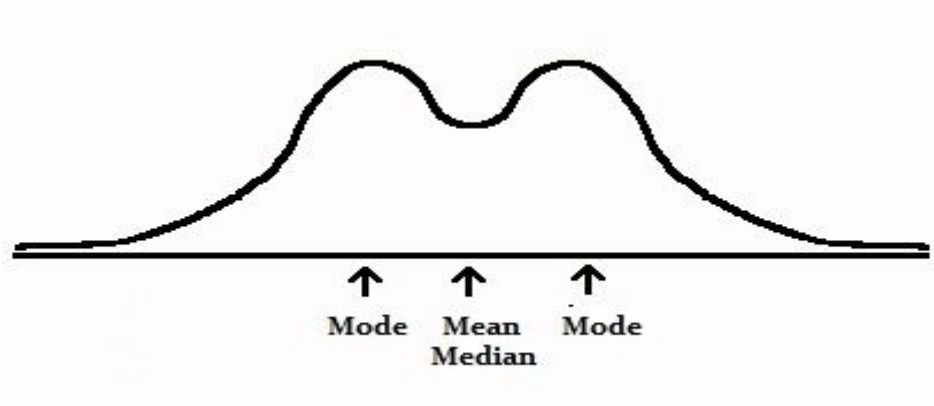
Today's discussion…

- 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

Symmetric Distribution

Distributions don’t have to be unimodal to be symmetric

The two halves are mirror images of each other



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

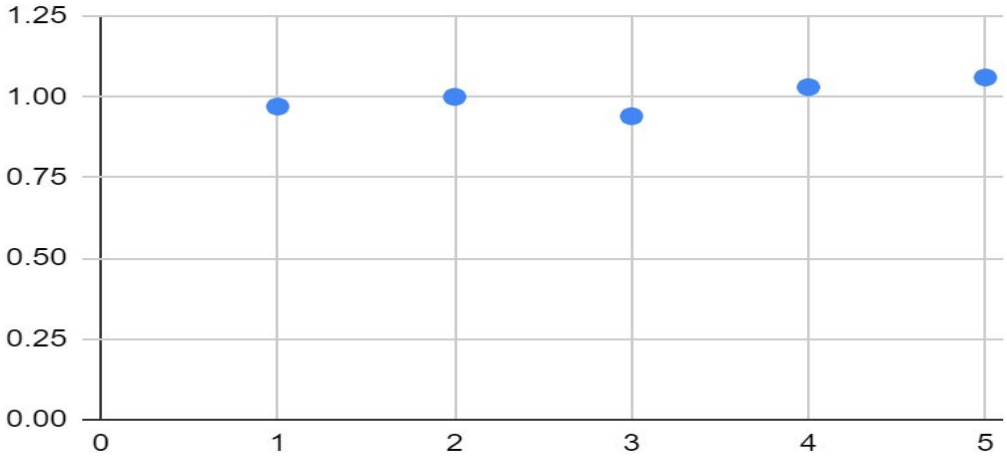
Example

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

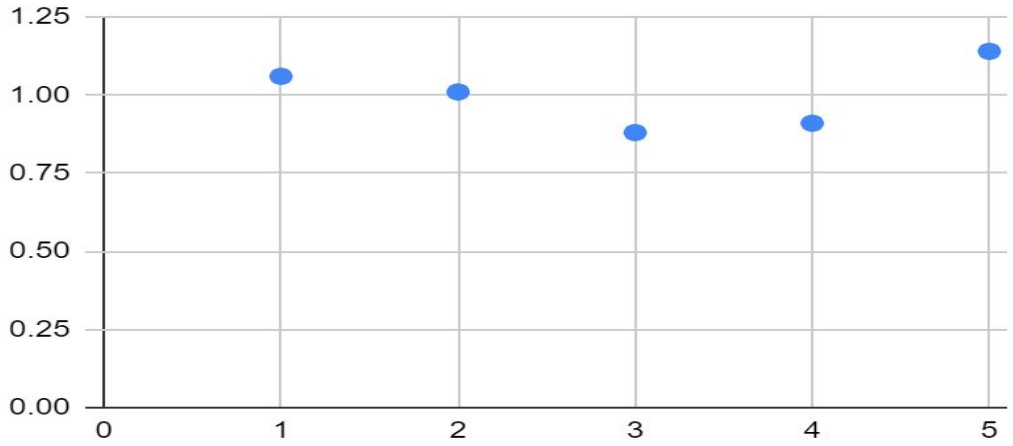
Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

- 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

Measures of dispersion



A



B

Range of a sample

Definition : **Range of a sample**

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

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

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

- 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 : **Variance and Standard Deviation**

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

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

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

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

Variance and Standard Deviation

- **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**
  - A low standard deviation- data observations tend to be very close to the mean
  - A high standard deviation- the data are spread out over a large range of values

Coefficient variation

Definition : **Coefficient 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.

Comparison of two data in terms of measures of central tendencies and dispersions in some cases will not be meaningful, because the variables in the data may not have same units of measurement.

Coefficient variation

The following table gives the values of mean and variance of heights and weights of the 10th standard students of a school.

	Height	Weight
Mean	155 cm	46.50 kg^2
Variance	72.25 cm^2	28.09 kg^2

Which is more varying than the other?

For Height

standard deviation $\sigma 1$ = 8. 5cm

C.V1= 5.48 %

For Weight

standard deviation $\sigma 2$ = 5.3kg

C.V2= 11.40 %

Since $C.V2 > C.V1$, the weight of the students is more varying than the height.

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

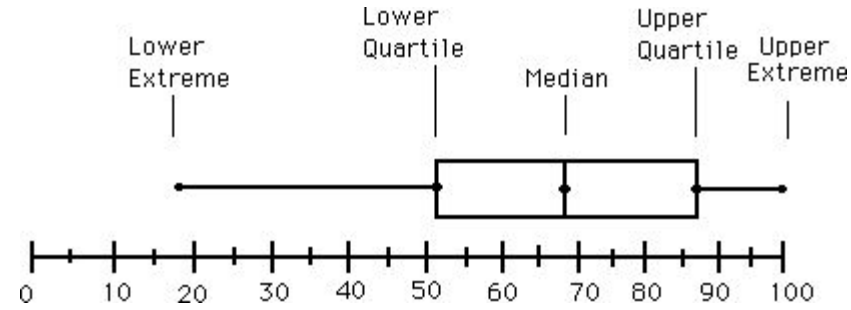
- **Five Number Summary**
  - Since, $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ together contain no information about the endpoints of the data, a complete summary of the shape of a distribution can be obtained by providing the lowest and highest data value as well. This is known as the five-number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

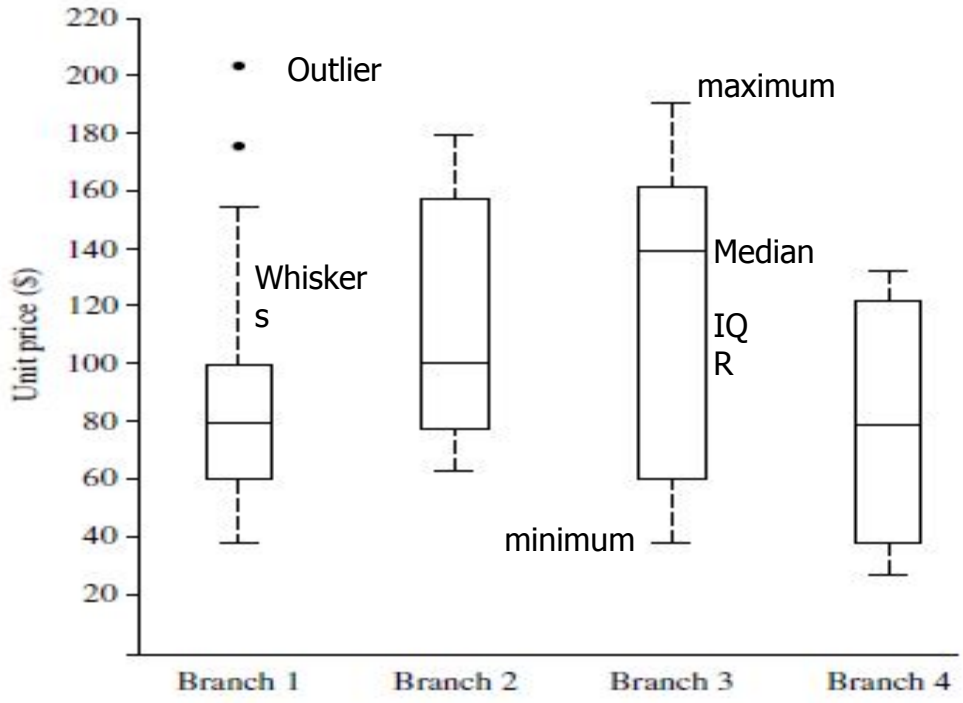
visualizing distribution



- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
    - The whiskers terminate at the most extreme observations occurring within 1.5**IQR of the quartiles*
  - Outliers: points beyond a specified outlier threshold, plotted individually

Box plot Analysis

- . *For branch 1, the median price of* items sold is 80, *Q1 is 60, and Q3 is 100. Notice two outlying observations for* this branch were plotted individually, as their values of 175 and 202 are more than 1.5 times the IQR .



Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

Example

Suppose that the data for analysis includes the attribute *age. The age values for the data* tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(d) Can you find (roughly) the first quartile (*Q1) and the third quartile (Q3) of the data?*

(e) Give the *five-number summary of the data.*

(f) Show a *boxplot of the data.*

Graphic Displays of Basic Statistical Descriptions

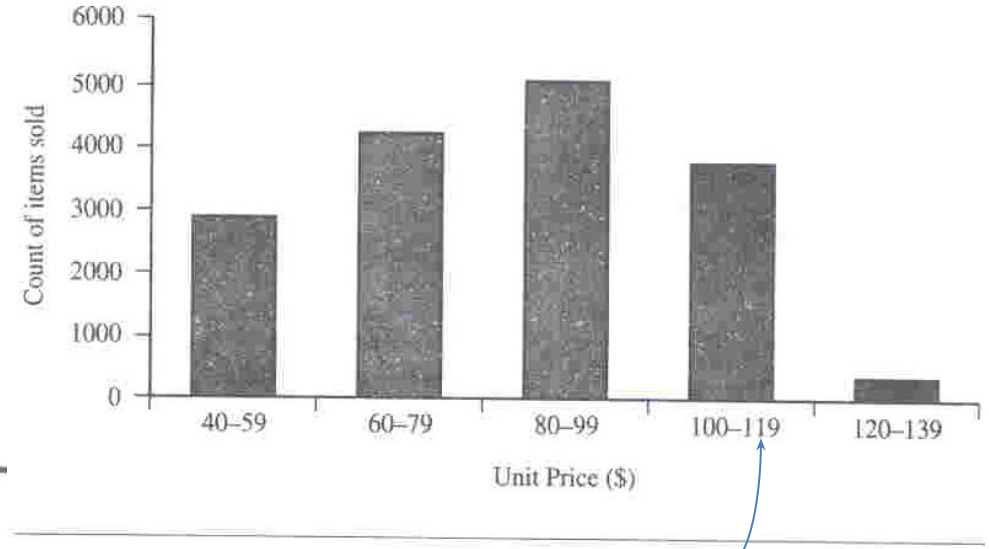
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Histogram Analysis

Summarizing the distribution of given attribute

A set of unit price data for items sold at a branch of *AllElectronics*.

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350

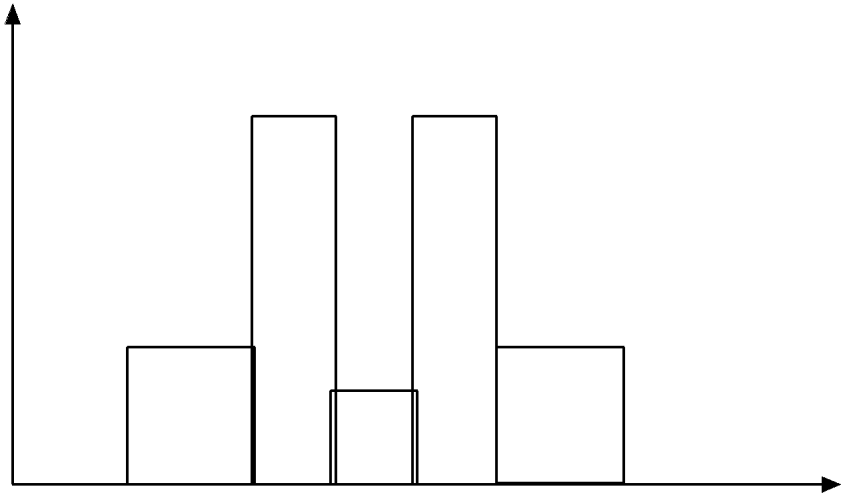


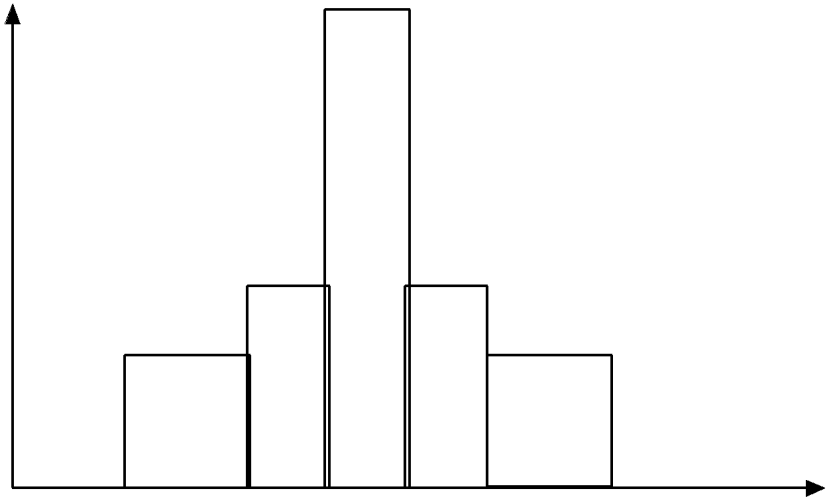
A histogram for the data set

Buckets or bins: disjoint subsets
which are uniform widthwise

Referred to as a bar chart if X is nominal(discrete)

Histograms Often Tell More than Boxplots





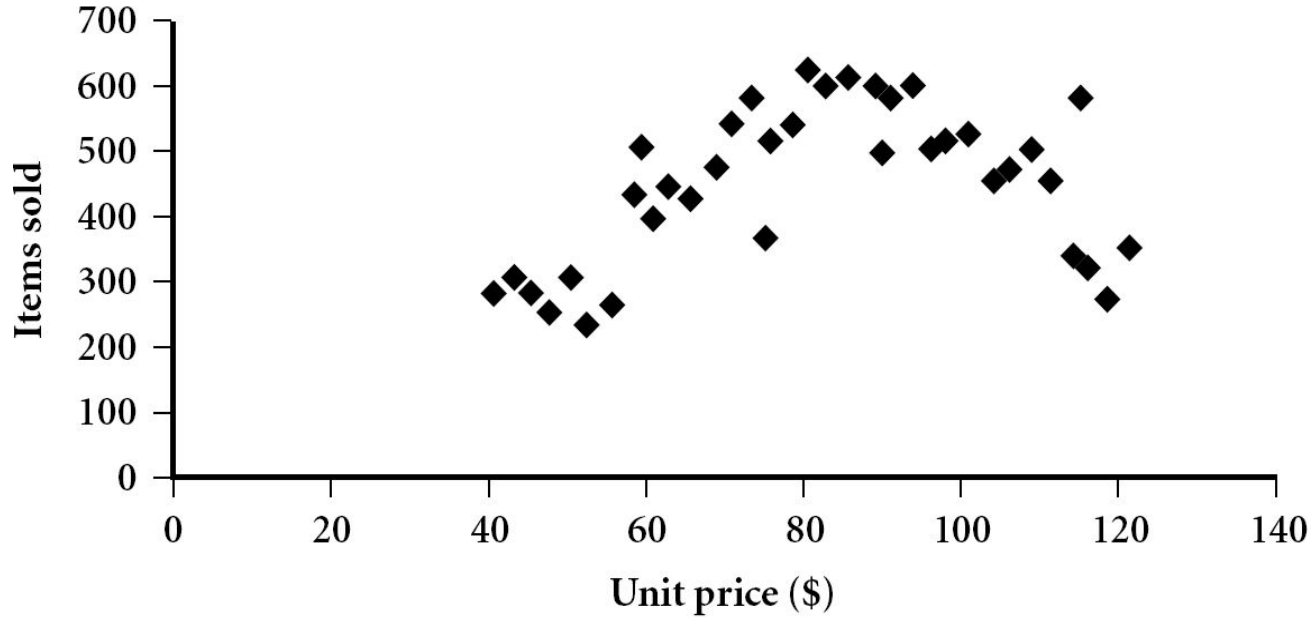
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Scatter Plots and Data Correlation

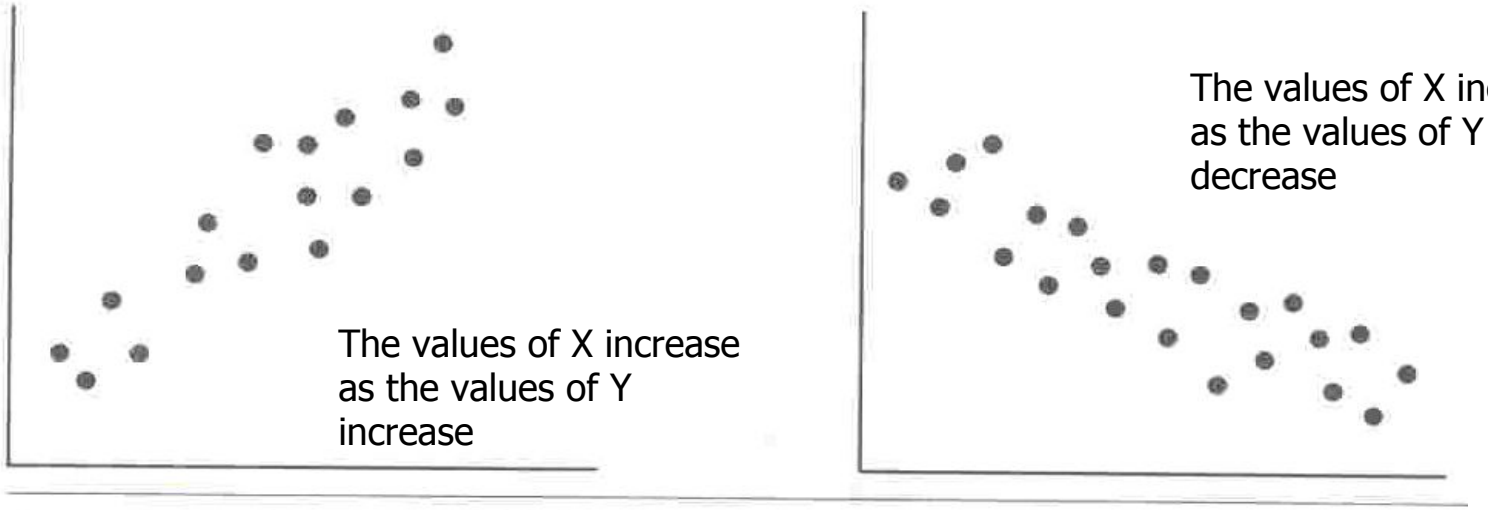
A set of unit price data for items sold at a branch of *AllElectronics*.

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350

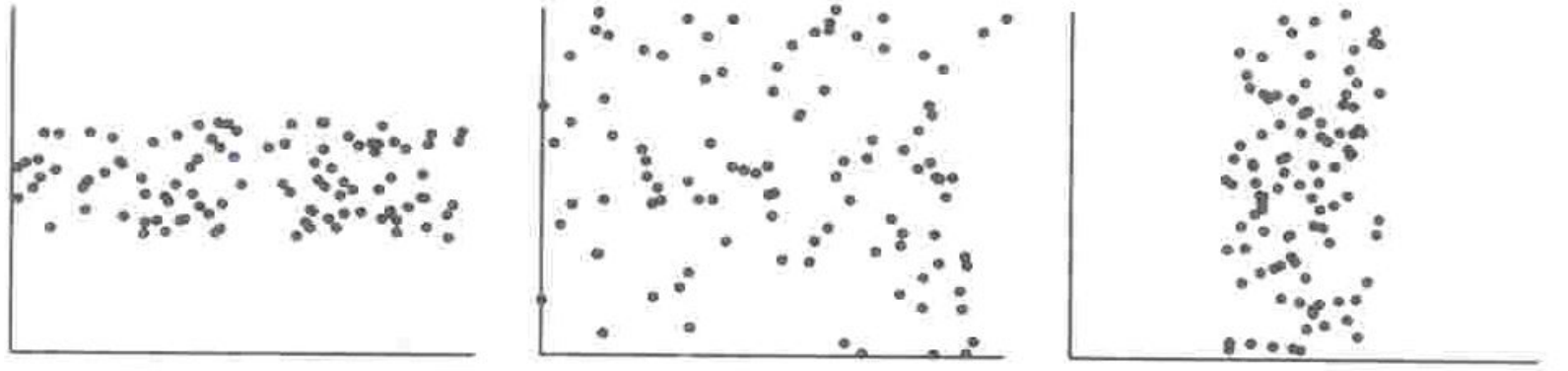
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Scatter Plots and Data Correlation



Scatter plots can be used to find (a) positive or (b) negative correlations between attributes

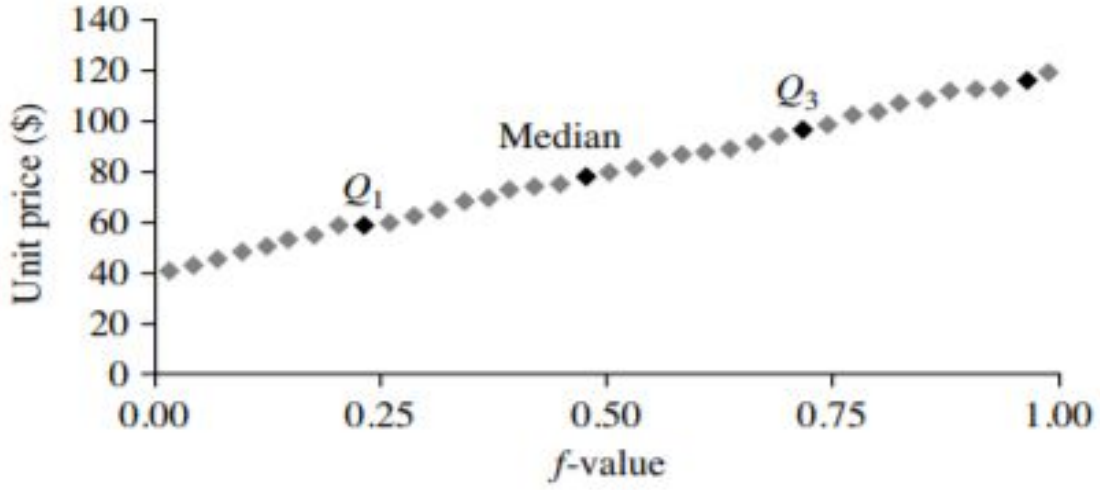


Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

Quantile Plot

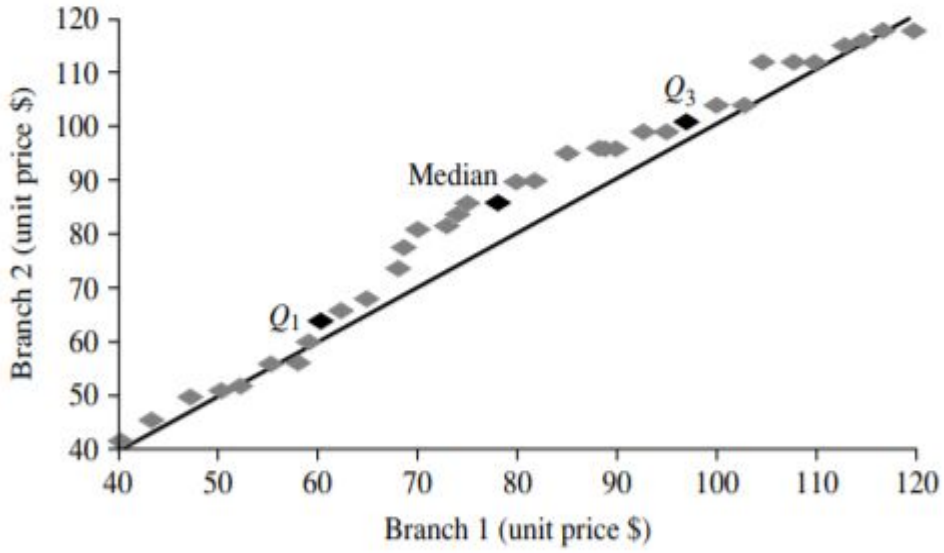
- Univariate data distribution
- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order
  - Each observation x_i is paired with a percentage f_i ,which indicates that approximately f_i * 100% of the data are below the value x_i
  - x_i is graphed against f_i

$$f_i = \frac{i - 0.5}{N}.$$



Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



50% of items sold at branch 1 were less than $78, while 50% of items at branch 2 were less than $85

If the interquartile range is zero, you can conclude that

(a) the range must also be zero

(b) the mean is also zero

(c) at least 50% of the observations have the same value

(d) all of the observations have the same value

Identify which of the following is a measure of dispersion

(a) median

(b) 90th percentile

(c) interquartile range

(d) mean

What is the primary characteristic of a set of data for which the standard deviation is zero?

(a) All values of the variable appear with equal frequency

(b) All values of the variable have the same value.

(c) The mean of the values is also zero.

(d) None of the above is correct.

The median is a better measure of central tendency than the mean if

(a) the variable is discrete

(b) the distribution is skewed

(c) the variable is continuous

(d) the distribution is symmetric

A sample of pounds lost in a given week by individual members of a weight reducing clinic produced the following statistics.

mean = 5 pounds	first quartile = 2 pounds
median = 7 pounds	third quartile = 8.5 pounds
mode = 4 pounds	standard deviation = 2 pounds

Identify the correct statement.

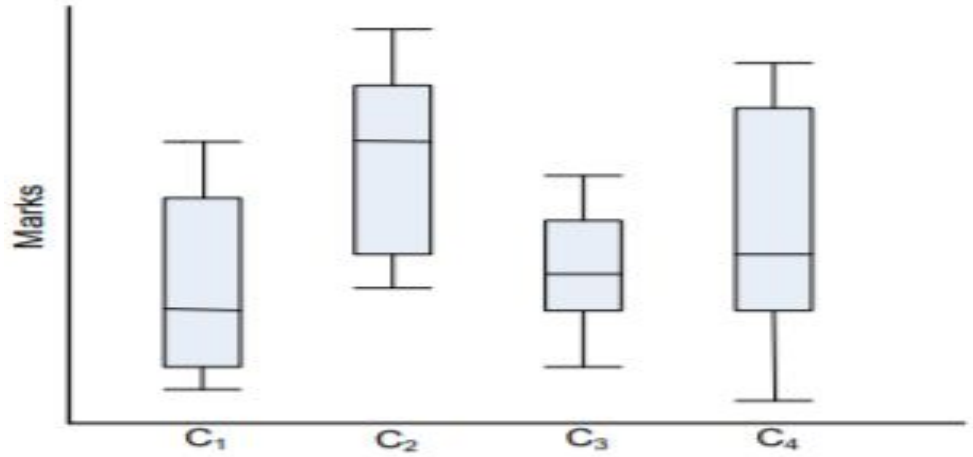
(a) One-fourth of the members lost less than 2 pounds.

(b) The middle 50% of the members lost between 2 and 8.5 pounds.

(c) The most common weight loss was 4 pounds.

(d) All of the above are correct.

From the tabulation of marks of students participated in four courses C1, C2, C3 and C4, box-plots are drawn, which is shown in Figure



The course in which students perform better is

(a) C1

(b) C2

(c) C3

(d) C4

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 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 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*?

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 vs. cumulative frequency)
    - More than (lower class vs. 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

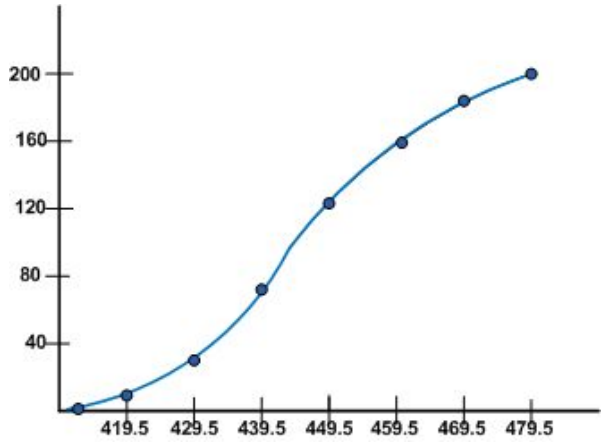
Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

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

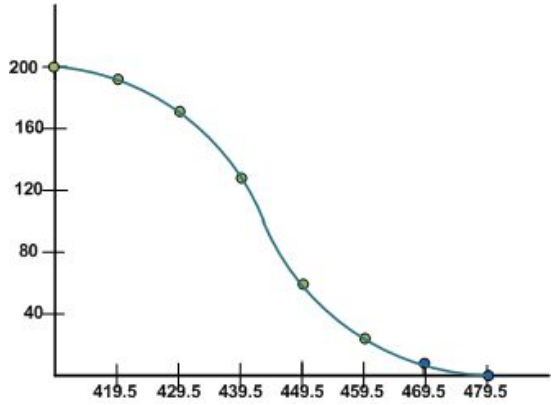


Ogive: Graphical method to find mean

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

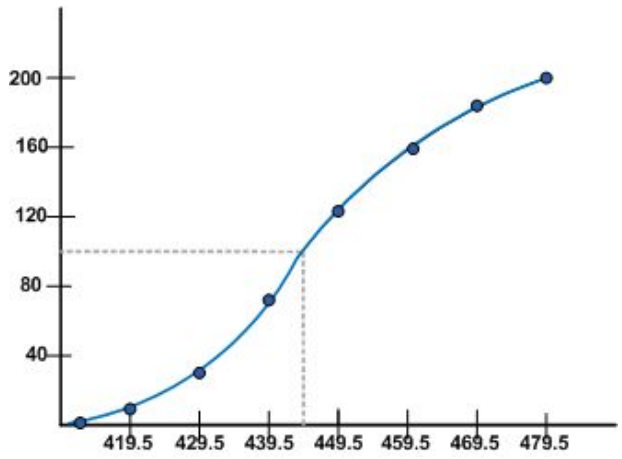
Step 3: More-than Ogive graph

Upper 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

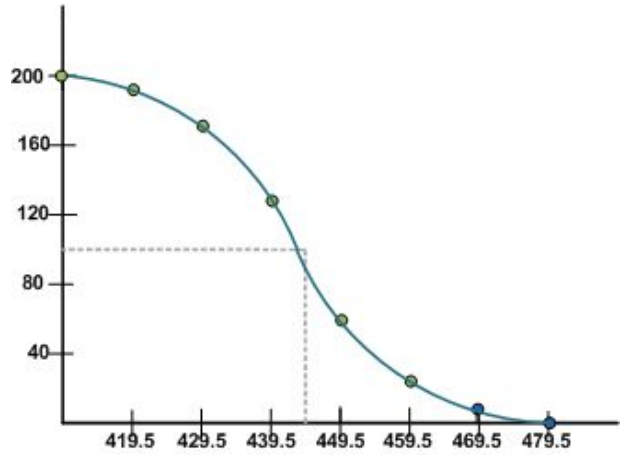


Information from Ogive

- Mean from Less-than Ogive



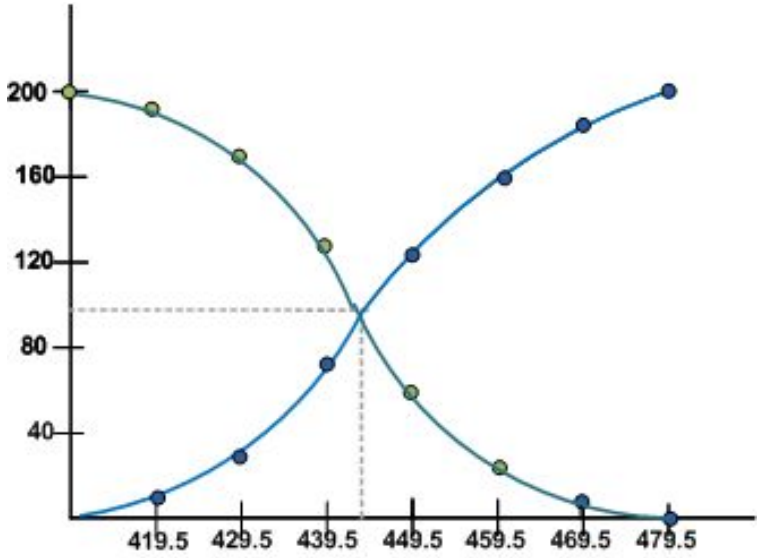
- Mean from More-than Ogive



To find the popularity of the given data or the likelihood of the data that fall within the certain frequency range,

Information from Ogive

- Less-than and more-than Ogive approach



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