Applied Statistics

Course Code: MAT1011

Dr. Sukanta Nayak

Department of Mathematics
School of Advanced Sciences
VIT-AP University, Amaravati
Andhra Pradesh

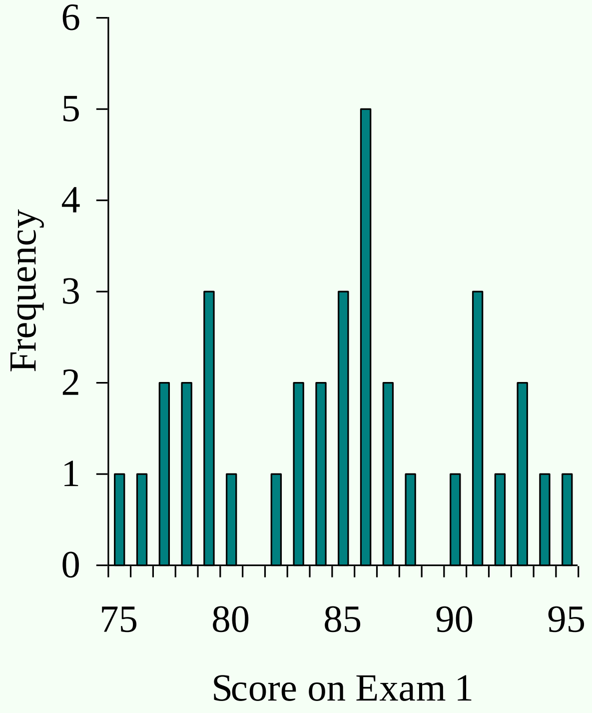
VIT-AP UNIVERSITY

Measures of Central Tendency

- A *measure of central tendency* is a descriptive statistic that describes the average, or typical value of a set of scores
- There are three common measures of central tendency:
  - the mode
  - the median
  - the mean

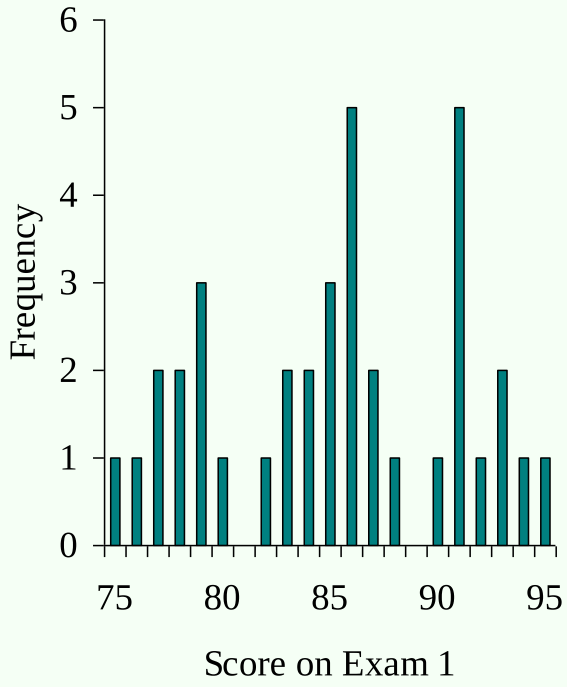
The Mode

- The *mode* is the score that occurs most frequently in a set of data



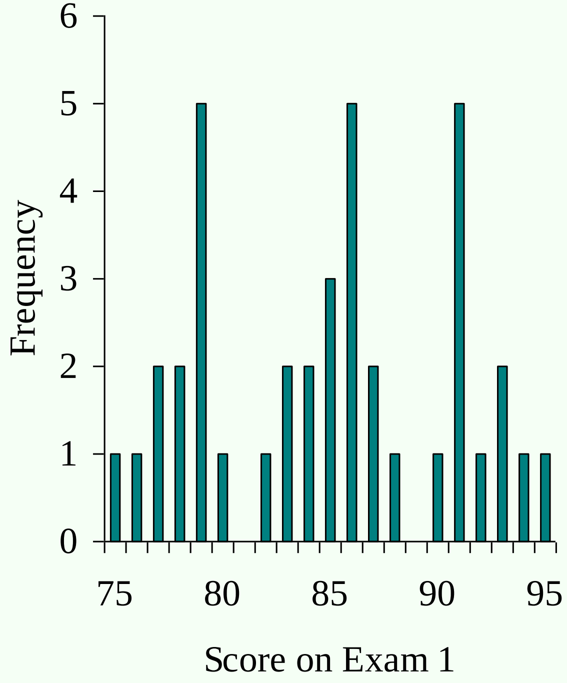
Bimodal Distributions

- When a distribution has two “modes,” it is called *bimodal*



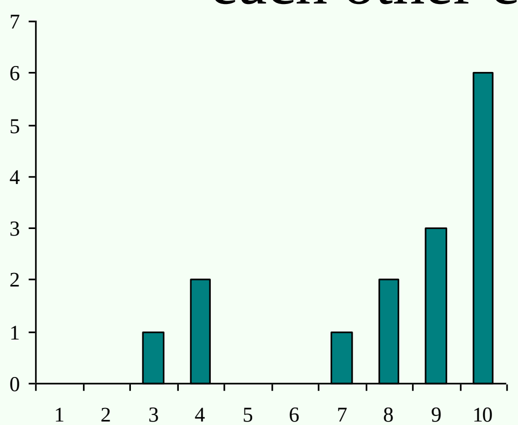
Multimodal Distributions

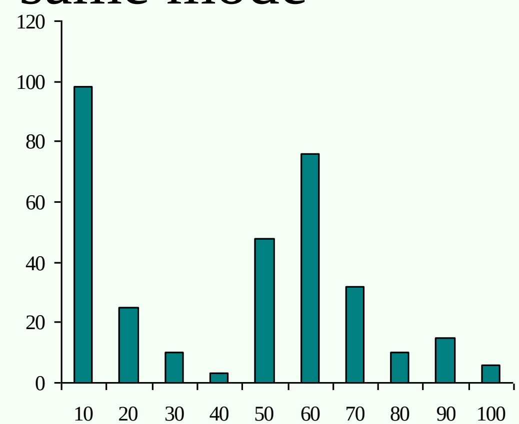
- If a distribution has more than 2 "modes," it is called *multimodal*



When To Use the Mode

- The mode is not a very useful measure of central tendency
  - It is insensitive to large changes in the data set
    - That is, two data sets that are very different from each other can have the same mode





When To Use the Mode

- The mode is primarily used with nominally scaled data
  - It is the only measure of central tendency that is appropriate for nominally scaled data

The Median

- The *median* is simply another name for the 50th percentile
  - It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median

How To Calculate the Median

- Conceptually, it is easy to calculate the median
  - There are many minor problems that can occur; it is best to let a computer do it
- Sort the data from highest to lowest
- Find the score in the middle
  - middle = (N + 1) / 2
  - If N, the number of scores, is even the median is the average of the middle two scores

Median Example

- What is the median of the following scores:
  10 8 14 15 7 3 3 8 12 10 9
- Sort the scores:
  15 14 12 10 10 9 8 8 7 3 3
- Determine the middle score:
  middle = (N + 1) / 2 = (11 + 1) / 2 = 6
- Middle score = median = 9

Median Example

- What is the median of the following scores:
  24 18 19 42 16 12
- Sort the scores:
  42 24 19 18 16 12
- Determine the middle score:
  middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5
- Median = average of 3rd and 4th scores:
  (19 + 18) / 2 = 18.5

When To Use the Median

- The median is often used when the distribution of scores is either positively or negatively skewed
  - The few really large scores (positively skewed) or really small scores (negatively skewed) will not overly influence the median

The Mean

- The *mean* is:
  - the arithmetic average of all the scores $(\Sigma X)/N$
  - the number, m, that makes $\Sigma(X - m)$ equal to 0
  - the number, m, that makes $\Sigma(X - m)^2$ a minimum
- The mean of a population is represented by the Greek letter μ; the mean of a sample is represented by X

Calculating the Mean

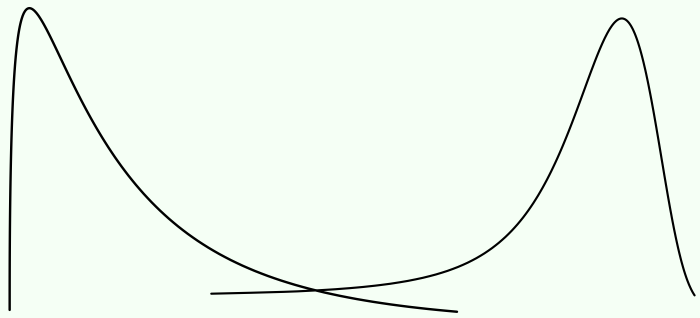
- Calculate the mean of the following data:
  1   5   4   3   2
- Sum the scores (ΣX):
  $1 + 5 + 4 + 3 + 2 = 15$
- Divide the sum ($\Sigma X = 15$) by the number of scores ($N = 5$):
  $15 / 5 = 3$
- Mean = X = 3

When To Use the Mean

- You should use the mean when
  - the data are interval or ratio scaled
    - Many people will use the mean with ordinally scaled data too
  - and the data are not skewed
- The mean is preferred because it is sensitive to every score
  - If you change one score in the data set, the mean will change

Relations Between the Measures of Central Tendency

- In symmetrical distributions, the median and mean are equal
  - For normal distributions, mean = median = mode
- In positively skewed distributions, the mean is greater than the median
- In negatively skewed distributions, the mean is smaller than the median



Mean of a sample

- The mean of a sample data is denoted as $\bar{\boldsymbol{x}}$. 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,….., 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 1: Simple Mean

If x_1, x_2, x_3,….., x_n are the sample values, the simple mean is defined as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_{\mathrm{i}}$$

Weighted mean of a sample

Weighted mean

It is also called weighted arithmetic mean or weighted average.

Definition 2: **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

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

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: **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,…,m* are the means of m samples of sizes n_1, n_2,….., 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 $\boldsymbol{x_k}$ is removed from a sample of size $\boldsymbol{n}$ with mean $\bar{\boldsymbol{x}}$, the new mean is given by

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

Lemma 4

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

$$\bar{\boldsymbol{x}} = \frac{\boldsymbol{n}\,\bar{\boldsymbol{x}}_{\boldsymbol{n}} \pm \boldsymbol{m}\,\bar{\boldsymbol{x}}_{\boldsymbol{m}}}{\boldsymbol{n} \pm \boldsymbol{m}}$$

Properties of mean

- **Lemma 5**

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

  $$\overline{\boldsymbol{x}}' = \overline{\boldsymbol{x}} \mp \boldsymbol{c}$$

- **Lemma 6**

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

  $$\overline{\boldsymbol{x}}' = \overline{\boldsymbol{x}} * \boldsymbol{c}$$

  *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, $x_i = \frac{1}{2}$ **(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

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

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)

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.*, 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 can assume that
  - » All classes are equal sized
  - » Groups are with inclusive classes, i.e., $\boldsymbol{x_i} = \boldsymbol{x_{i-1}}$ (*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 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, …….

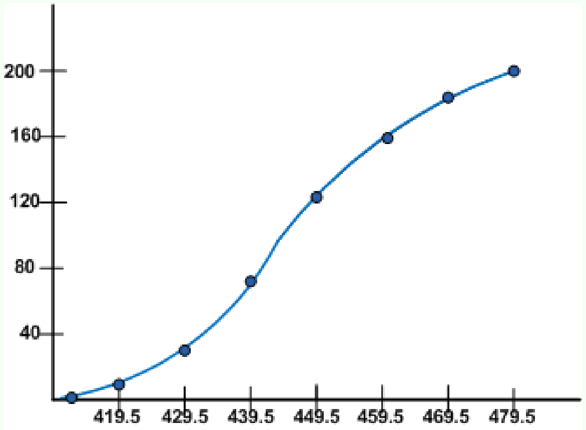
Step 1: Draw a cumulative frequency table

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

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



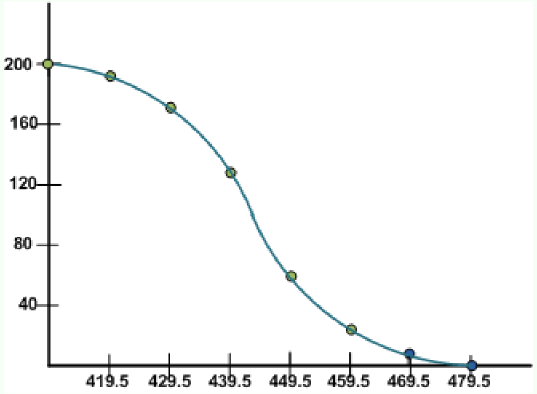
Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

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

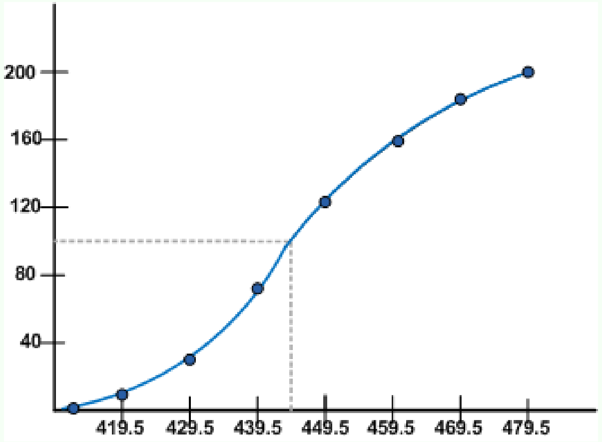
Step 3: More-than Ogive graph

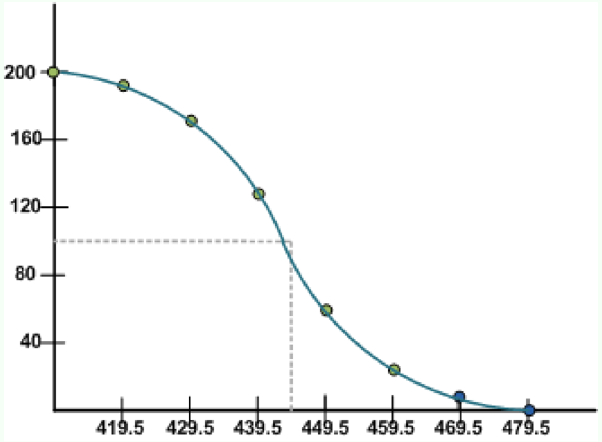
Upper class	Cumulative Frequency
More than 409.5	200
More than 419.5	186
More than 429.5	166
More than 439.5	124
More than 449.5	70
More than 459.5	25
More than 469.5	7



Information from Ogive

- Mean from Less-than Ogive
- Mean from More-than Ogive

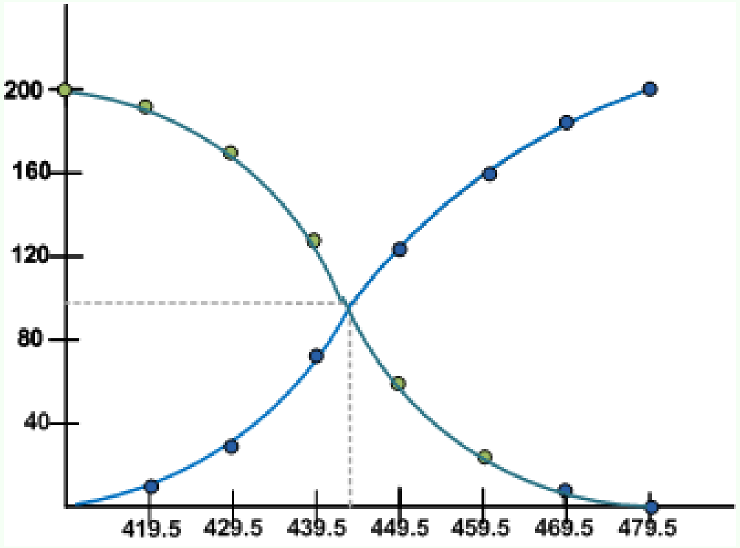




- A % C frequency of .65 for the third class 439.5.....449.5 means that 65% of all scores are found in this class or below.

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)

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**)
  - $S: \{x_1, x_2\}$
  - $\hat{x} = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$
  - $\frac{2}{\hat{x}} = \frac{1}{x_1} + \frac{1}{x_2}$

Questions

» Is there any generalization for AM ($\bar{\boldsymbol{x}}$), GM ($\tilde{\boldsymbol{x}}$) and HM ($\hat{\boldsymbol{x}}$) 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 5: Geometric mean

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

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

where, $n \neq 0$

Note

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

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

» 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 6: **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, $\boldsymbol{\hat{x} = e^{\bar{y}}}$
- 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, $\boldsymbol{\tilde{x} = \frac{1}{\bar{y}}}$

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 7: **Median of a sample**

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

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

Median of a sample

Definition 8: **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 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 9: **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

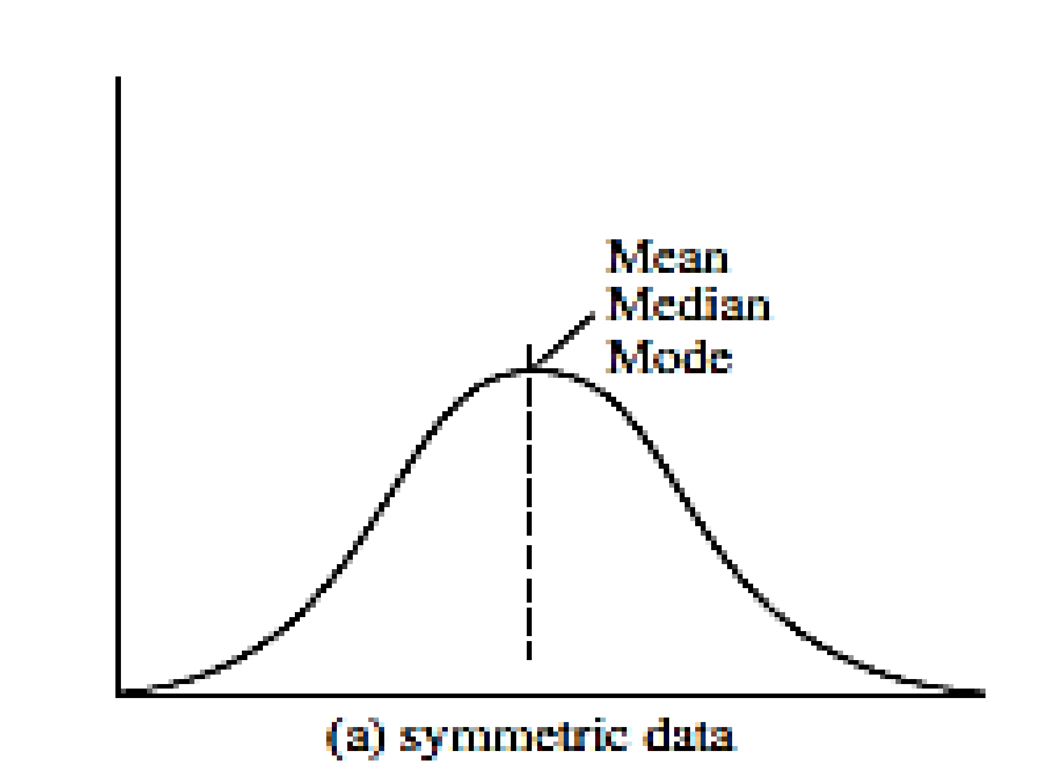
- To understand the above three categories, let us consider the following
- Given a set of $\boldsymbol{m}$ objects, where any object can take values $\boldsymbol{v_1}$, $\boldsymbol{v_2}$,…..,$\boldsymbol{v_k}$. Then, the frequency of a value $\boldsymbol{v_i}$ is defined as

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

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

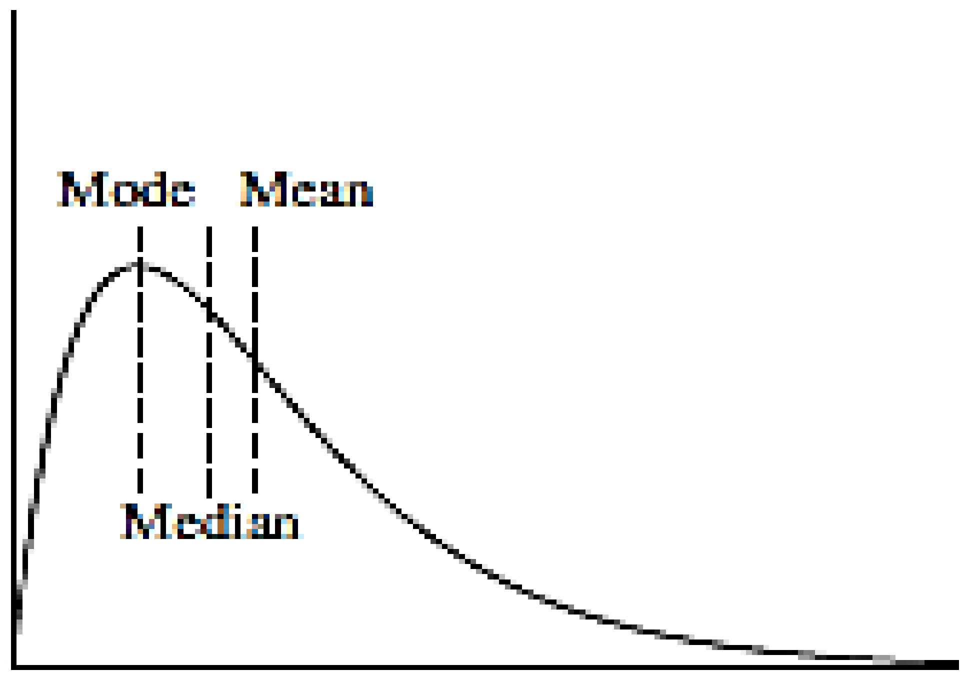
Symmetric data

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



Positively skewed data

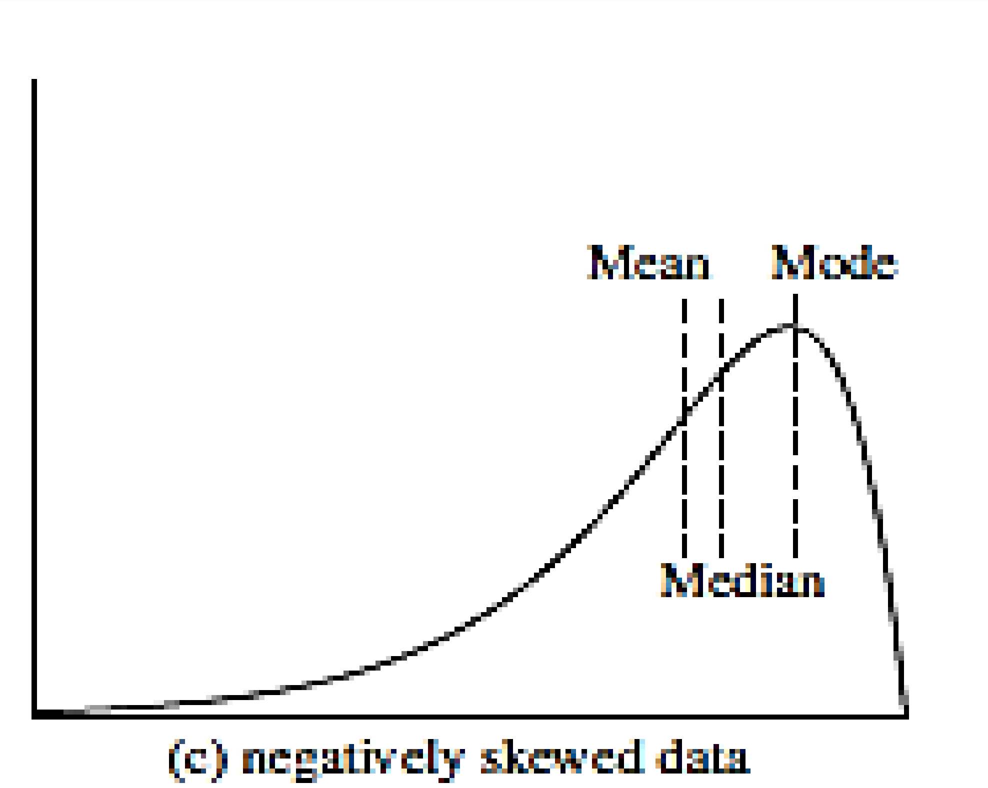
- Here, mode occurs at a value smaller than the median



(b) positively skewed data

Negatively skewed data

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



Empirical Relation!

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

$$\text{Mean} - \text{Mode} = 3 * (\text{Mean} - \text{Median})$$

References

1. Norean R. Sharpe, Richard D. De Veaux, Paul F. Velleman, Business Statistics, Fourth Edition, Pearson Education. 2019.
2. R.E. Walpole, R.H. Mayers, S.L. Mayers and K.Ye, Probability and Statistics for engineers and scientists, 9th Edition, Pearson Education, 2018.
3. S. C. Gupta and V. K. Kapoor, Fundamentals of Applied Statistics, S. Chand, 2006.
4. Douglas C. Montgomery, George C. Runger, Applied Statistics and Probability for Engineers, Sixth Edition, John Wiley and Sons, 2014.
5. Miller & Freund’s, Probability and statistics for engineers, 8th edition, Pearson publication, 2018.

THE END