Measures of Central Tendency

Dr. Umesh R A

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Agenda

Measures of Central Tendency

- Arithmetic Mean
- Geometric Mean
- Harmonic Mean
- Mode
- Median

Partition Values

- Quartile
- Decile
- Percentile
- Quantile

What is Measure of Central Tendency?

What is Measure of Central Tendency?

- **Central Tendency:**

Property of concentration of the observations around central value is called Central Tendency.

- **Measures of Central Tendency:**

The central value around which there is concentration is called Measures of Central Tendency.

What are the Measures of Central Tendency?

Measure of Central Tendency

- Arithmetic Mean (or simply Mean)
- Median
- Mode
- Harmonic Mean (H.M.)
- Geometric Mean (G.M.)

What are the criteria's to judge a goodness of Measure of Central Tendency?

Requisites of Good Averages

- It should be based on all the observations.
- It should not be affected by extreme value.
- It should be rigidly defined.
- It should be easy to calculate and understand.
- It should be capable of further algebraic treatment.
- It should work in case of open end class intervals.

Arithmetic Mean (AM)

"AM is obtained by dividing sum of all the values by the total number of observations"

$$\bar{x} = \frac{Sum\ of\ all\ Observations}{Total\ number\ of\ observations} = \frac{\sum x}{n}$$

- Arithmetic mean is a mathematical average.
- It is frequently referred to as 'Mean'.
- It is denoted by $\overline{\text{x}}$.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Arithmetic Mean: Formulae

For Raw Data:

$$\bar{x} = \frac{Sum\ of\ all\ Observations}{Total\ number\ of\ observations} = \frac{\sum x}{n}$$

For Grouped Data:

$$\bar{x} = \frac{\sum fx}{N} = \frac{\sum fx}{\sum f}$$

Properties of A.M.

1. AM is not independent of change of origin and scale.

2. The sum of the deviations of the observations from the AM is always zero.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

[cont…]

3. The sum of the squared deviations of the observations from the AM is a always minimum (called as Minimal property of AM).

$$\sum(x-\bar{x})^2 < \sum(x-a)^2$$

4. **Combined Mean:** we can obtain combined mean of two different sets of observations using following formula.

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Shortcut method

Change of origin and scale method:

$$\bar{x} = A + \left[B \times \frac{\sum \mathrm{fu}}{N} \right] when, u = \frac{x - A}{B}$$

Weighted Arithmetic Mean

Simple AM gives equal importance to all observations.

But, sometimes all values may have unequal importance.

We calculate Weighted Arithmetic Mean.

$$\bar{x} = \frac{\sum wx}{\sum w}$$

Weighted Arithmetic Mean

Example:

If we want to find the **average marks per student** in different subjects like **Mathematics, Statistics, Physics** and **Biology**. These subjects do not have equal importance while admitting students for Data Science course.

Advantages of Mean

- It is easy to understand & simple calculate.
- It is based on all the values.
- It is rigidly defined.
- It can be useful when data consist one or more zero or negative values.
- It is not based on the position in the series.
- It is least affected by fluctuations of sampling.

Disadvantages of Mean

- It is affected by abnormal extreme values.
- It cannot be calculated for open end classes.
- Not suitable for categorical data.
- It cannot be computed if any value is missing.
- The mean sometimes does not coincide with any of the observed value. Sometimes AM may have unrealistic value.

Geometric Mean (GM)

"Geometric Mean of n observations is the nth root of the product of the given observations"

$$\mathrm{G.M.}\,(G) = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$$

$$strictly\ x \geq 0.$$

- Geometric mean is a mathematical average.
- It is denoted by G.

Geometric Mean: Formulae

Simplified Formula for Raw Data:

$$G.M. = \text{Antilog}\left[\frac{1}{n}\sum \log(x_i)\right]$$

Simplified Formula for Grouped Data:

$$G.M. = \text{Antilog}\left[\frac{1}{\sum f_i}\sum f_i \log(x_i)\right]$$

Advantages of Geometric Mean

- Geometric mean is based on all the observations.
- It is useful for further algebraic treatment.
- It is not affected by sampling fluctuations.
- It gives the greatest weight to the smallest item in the data.
- It is suitable if the distribution is skewed.

Disadvantages of Geometric Mean

- It cannot be calculated if any value is zero.
- It cannot be calculated if any value is missing.
- It can not be calculated when data contains any negative value.
- Geometric mean is moderately affected by the abnormal extreme value.
- It is usually a value which does not exist in the given data.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Harmonic Mean (HM)

"Harmonic Mean of 'n' observations is the reciprocal of the arithmetic mean of the reciprocals of all observations"

$$\text{H.M.}\,(H) = \frac{n}{\sum \frac{1}{x}}; \qquad strictly\; x \neq 0.$$

Harmonic mean is a mathematical average.

It is denoted by H.

Harmonic Mean: Formulae

For Raw Data:

$$\text{H. M.}\,(H) = \frac{n}{\sum \frac{1}{x}}; \quad strictly\ x \neq 0.$$

For Grouped Data:

$$\text{H. M.}\,(H) = \frac{N}{\sum \frac{f}{x}} \quad OR \quad \frac{\sum f}{\sum \frac{f}{x}}; \quad strictly\ x \neq 0.$$

Advantages of Harmonic Mean

- It is based on all the observations.
- Useful for further algebraic treatment.
- It is not affected by sampling fluctuations.
- It gives the greatest weight to the smallest value.
- It can be calculated even when a series contains any negative value.
- It is suitable if the distribution is skewed.

Disadvantages of Harmonic Mean

- Harmonic mean cannot be calculated if any value is missing.
- Harmonic mean is slightly affected by the abnormal extreme value.
- It is usually a value which does not exist in the given data.

Median (M)

"It is a middle most value, when data is arranged in ascending or descending order."

- Median is the value which divides the data into two equal parts.
- Median is a positional average.
- It is denoted by M.

Median: Formulae

Raw data:

Step 1. Arrange data in Ascending or descending order.

Step 2. Apply following formula;

$$Median\ (M) = \left(\frac{\mathrm{n}+1}{2}\right)^{th}\ obeservation$$

'n' is total number of observations

Discrete Frequency Distribution:

Step 1. Calculate Less than Cumulative Frequencies (LCF).

Step 2. Apply following formula;

$$Median\ (M) = \left(\frac{N+1}{2}\right)^{th}\ obeservation$$

'N' is total number of observations $= \sum f$

Continuous Frequency Distribution:

Step 1. Calculate LCF.

Step 2. Find median class = $\left({}^{N}/_{2}\right)^{th}$ obs. containing class.

Step 3. Apply following formula;

$$Median\ (M) = l_1 + \frac{\left(\frac{N}{2} - pcf\right) * c}{f}$$

l_1	: lower limit of median class
pcf	: previous cumulative freq. of median class
c	: class width
f	: class frequency
N	: total number of observations $(\sum f)$

Example:

Find Median Age from the following data;

Age Group	Frequency (f)	LCF
0-20	15	15
20-40	32	47
40-60	54	101
60-80	30	131
80-100	19	150
Total	**150**	

Solution:

$$\text{Median (M)} = 40 + \frac{\frac{150}{2} - 47}{54} \times 20$$

$$= 40 + \frac{75 - 47}{54} \times 20$$

$$= 40 + \frac{28}{54} X 20$$

$$= 0 + 0.52X20$$

$$= 40 + 10.37$$

$$= \mathbf{50.37}$$

Aegis

SCHOOL OF BUSINESS

SCHOOL OF DATA SCIENCE

SCHOOL OF TELECOMMUNICATION

Property of Median

Minimal Property of Median:

The sum of the absolute values of deviation of the item from Median is always minimum.

i.e. $\sum|(x-M)| < |(x-a)|$

Here, '*a*' is any value other than median

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Advantages of Median

- Median can be calculated in all distributions.
- It does not affected by abnormal extreme values.
- Especially useful in case of open-end classes.
- It can be located graphically (using Ogive’s).
- It is most useful dealing with <u>ordered</u> qualitative data.

Disadvantages of Median

- It is not based on all the values.
- It is not capable of further mathematical treatment.
- It is affected fluctuation of sampling.
- In case of even no. of values it may not the value from the data.

Mode (Z)

"Mode is the most frequent (most frequent) value in the distribution"

$$\text{Mode}\ (Z) = Most\ frequent\ obeservation$$

- Mode is most times repeated value.
- Mode is a positional average.
- It is denoted by Z.

Mode: Formulae

Raw data:

Step 1. Construct frequency distribution.

Step 2. Apply following formula;

$$\text{Mode } (Z) = Most\ frequent\ obeservation$$

SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

[Cont…]

Discrete Frequency Distribution:

Apply following formula;

$$\text{Mode}\ (Z) = Most\ frequent\ obeservation$$

[Cont…]

Continuous Frequency Distribution:

Step 1. Find modal class (highest freq. class).

Step 2. Apply following formula;

$$Mode\ (Z) = l_1 + \frac{(f_1 - f_0) * c}{(2f_1 - f_0 - f_2)}$$

l_1	: lower limit of modal class
f_1	: frequency of modal class
f_0	: frequency of <u>preceding</u> class to modal class
f_2	: frequency of <u>next</u> class to modal class
c	: class width

Example:

Find Mode for monthly rent Paid by XYZ bank ATM in Kerala;

Monthly rent (Rs)	Number of Libraries (f)
500-1000	5
1000-1500	10
1500-2000	8
2000-2500	16
2500-3000	14
3000 & Above	12
Total	**65**

Solution:

$$Z = 2000 + \frac{16-8}{2(16)-8-14} \times 500$$

$$Z = 2000 + \frac{8}{32 - 8 - 14} X500$$

$$Z = 2000 + \frac{8}{10} \times 500$$

Z=2000+0.8 ×500=400

Z=**2400**

Advantages of Mode

- It is not affected by extreme value.
- The mode can be determined graphically (using Histogram).
- It is usually an actual value of an important part of the series.
- Even when some of the extreme values are missing , it can be computed.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Disadvantages of Mode

- It is not based on all observations.
- It is not capable of further algebraic treatment.
- For some frequency distributions Mode is ill-defined.
- Choice of grouping has great influence on the value of mode.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Which one is better?

AM or GM or HM or Median or mode?

'Share your opinion…'

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Discussion!

Not all data is the same.

The proper selection of a measure is decided according to;

- Symmetry of the distribution
- Scale of Measurement
- Qualitative vs Quantitative
- Purpose of research (objective)
- Outliers presence

Symmetry of the distribution

If you have interval / ratio data, but the distribution is very asymmetric, then the median is usually a better choice than the mean.

Example: Income is an extremely skewed variable for which the median is usually considered as suitable measure.

Scale of measurement

Interval/ratio (not skewed): Mean
Interval/ratio (skewed) : Median, Mode
Ordinal : Median
Nominal : Mode

Quantitative vs Qualitative data

- Quantitative data: It depends on scale of measurement

- Qualitative data:

  -for ordered data : Median

  -for unordered data : Mode

Purpose of the research

Purpose of the research (Research Objective)

- Most typical value : Mode
- Middle value : Median
- Average value : Mean

Outliers

If outliers are present in the data set then Median is suitable measure over mean.

How to decide between AM, GM and HM?

Here are some tips!

- Use arithmetic mean, when data varies in the same interval (no outliers).
- Use geometric mean, when data contains fractions, growth rates etc.
- Use harmonic mean, when data contains fractions and/or extreme values (either too big or too small).

In simple words;

- Arithmetic mean when numbers are additive in nature.
- Geometric mean when numbers are multiplicative in nature.
- Harmonic mean when the reciprocals of the numbers are additive in nature.

…Lets see couple of trivial example…

Geometric Mean

Example:

Situation 1: The profit of Company A, XYZ Ltd., has grown over the last three years by 10 million, 12 million, and 14 million dollars. It is appropriate to say that it has grown by an average of 12 million dollars yearly, for which we use the arithmetic mean. *(Suitable Measure is AM)*

Situation 2: The profit of Company B, XYZ Ltd., has grown the over last three years by 2.5%, 3%, and 3.5%. *(Here we cannot use the Arithmetic Mean and say that the average growth was 3%. Why not?)*

Explanation

In case of Arithmetic Mean calculation will be;

Suppose that Company B, XYZ Ltd., started with a 100-million-dollar profit. Three years later it will have become:

$100,000,000 * 1.025 * 1.03 * 1.035 = **$109,270,125**

Lets calculate Arithmetic mean of growth rates,

AM= (1.025 + 1.03 + 1.035) / 3 =1.03

Actual value is less than a yearly increase of 3% would yield, since:

$100,000,000 * 1.03 * 1.03 * 1.03 = **$109,272,700**

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Explanation

<u>In case of Geometric Mean calculation will be;</u>

Suppose that Company B, XYZ Ltd., started with a 100-million-dollar profit. Three years later it will have become:

$100,000,000 * 1.025 * 1.03 * 1.035 = **$109,270,125**

Lets calculate Geometric mean of growth rates,

GM=(1.025 * 1.03 * 1.035)^(1/3) = 1.02999190932156

Observe!! geometric mean result will be equal to actual value:

$100,000,000 * 1.02999190932156 * 1.02999190932156 * 1.02999190932156 = **$109,270,125**

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

GM vs AM

Therefore,

- Arithmetic mean when numbers are additive in nature.
- Geometric mean when numbers are multiplicative in nature.

Harmonic Mean

- This amounts to the reciprocal of the arithmetic mean of the RECIPROCALS of the individual speeds.

- In general, we use the harmonic mean when the numbers naturally combine via their reciprocals.

Harmonic Mean

- **Example:** If you want to calculate average speed of a car that goes the **same distance with different speeds**, then the net effect of all the driving is found by dividing the common distance by each speed to get the time for that trip.
- In general, we use the harmonic mean when the numbers naturally combine via their reciprocals.

Hence,

- Arithmetic mean when numbers are additive in nature.
- Geometric mean when numbers are multiplicative in nature.
- Harmonic mean when the reciprocals of the numbers are additive in nature.

Partition Values

What is partition values?

What are the partition values?

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Partition Values

Partition values divides distribution (data) in equal ‘k’ parts.

Partition values are;

- Median
- Quartiles
- Deciles
- Percentiles
- Quantiles

Median

Median is one of the partition value which divides data into equal 2 parts.

3 5 5 **7** 9 11 15

⌂

Median

Quartiles

Quartiles divides a data into 4 equal parts. For any series there are three quartiles denoted by Q_1,Q_2 and Q_3.

Q_1: first or lower quartile, covering 25% items.

Q_2: second quartile or median, covering 50%.

Q_3: third or upper quartile, covering 75% items.

Quartiles

	Q_1		Q_2		Q_3	
First Quarter		Second Quarter		Third Quarter		Fourth Quarter
24 25 26		27 30 32		40 44 50		52 55 60
	26.5		**31.0**		**51.0**	

Deciles

Deciles divides a data into 10 equal parts.

For any series, there are 9 deciles denoted by $D_1, D_2, \ldots, D_9$.

These are called as 1^{st} Decile, 2^{nd} Decile and so on.

Percentiles

Percentiles divide a data into 100 equal parts.

For any series, there are 99 percentiles denoted by $P_1, P_2, \ldots, P_{99}$.

Quantiles

- Its just generic term.

- Quantile divides data into equal 'k' groups, each containing the same fraction of the total population.

- Example: quartile, percentile, etc.

Formulae: Discrete Distribution

Note: First find Cumulative Frequency (LCF)

$Q_i = \{i(N+1)/4\}^{th}$ observation

where, i=1,2,3

$D_i = \{i(N+1)/10\}^{th}$ observation

where, i=1,2,...9

$P_i = \{i(N+1)/100\}^{th}$ observation

where, i=1,2,...,99

Formulae: Continuous Distribution

1. First find Cumulative Frequency (LCF)
2. Then, find Qi or Di or Pi

$$Q_{\mathrm{i}} = l + \frac{\left(\frac{iN}{4} - pcf\right) \cdot c}{f} \quad \textit{where, i=1,2,3}$$

$$D_{\mathrm{i}} = l + \frac{\left(\frac{iN}{10} - pcf\right) \cdot c}{f} \quad \textit{where, i=1,2,...9}$$

$$P_{\mathrm{i}} = l + \frac{\left(\frac{iN}{100} - pcf\right) \cdot c}{f} \quad \textit{where, i=1,2,...,99}$$

Next Lecture

- Topic: **Descriptive Statistics II**
  - **Measures of Dispersion (Range, MD, SD, Variance, CV)**
  - **Skewness, Kurtosis**
- Where you will find reference to study
  - Book Statistics Class-11, Chapter1, pp.81-165
- Background material to study
  - Business Statistics, Chpt1, pp.10-41 & Chpt5, pp.205-207
- We will have MCQ on this lecture.
- And will have recap of the lecture one.
- Discuss on the assignment of this lecture.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION