Descriptive Statistics II

Dr. Umesh R A

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF CYBER SECURITY
SCHOOL OF TELECOMMUNICATION

Agenda

- Measures of Dispersion
- Skewness
- Kurtosis

Is Central Tendency of data is enough to describe the frequency distribution?

Yes or No?

What is Measures of Dispersion (Variation)?

Why Measures of Dispersion?

Dispersion

Dispersion:

The property of deviation of values from the average is called variation or dispersion.

Measures of Variations:

The degree of variation is indicated by Measures of Variations.

Dispersion

Measures of Variation:

- Range
- Interquartile Range
- Mean Deviation (M.D.)
- Standard Deviation (S.D.)

The above measures are absolute measures of variation.

Relative measures of variation

However, for comparison of variation in two or more frequency distribution, a relative measure of variation should be calculated.

Relative measures of variation:

1. Coefficient of Range
2. Coefficient of Quartile Deviation
3. Coefficient of Mean Deviation
4. Coefficient of Variation (C.V.)

Absolute vs Relative Measures

- The **absolute measures are dependent of the units of Measurement**. So, we can not compare two different frequency distributions directly.

- The **relative measures are independent of the units of Measurement**. And so, they **facilitate comparison**.

Requisites of Good Measure of Dispersion

- 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 be Possible to find for open end class intervals.

Range

Range is the difference between the highest and the lowest value in the data.

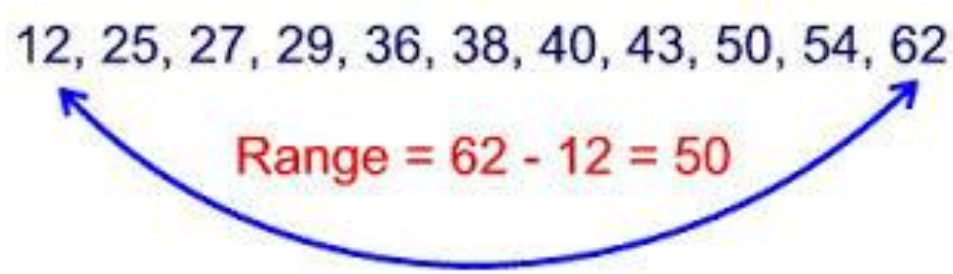
$$R = H - L$$

where, H-Highest Value, L-Lowest Value

Coefficient of range:

$$Coef.of\ Range = \frac{H - L}{H + L}$$

Example:



Range

- A major drawback of Range is that, since it is based on extreme values, it is highly affected by abnormal values.

- This drawback is rectified in quartile deviation.

Quartile Deviation (Q.D.)

Quartile Deviation (or Semi-Interquartile Range)

- It is based on only lower and upper Quartiles.
- It is not based on all the observations.
- So, it is not affected by abnormal extreme values.

Q.D.: Formulae

Quartile Deviation:

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

- Coefficient of Quartile deviation:

$$Coef.Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

- Interquartile Range (IQR):

$$IQR = Q_3 - Q_1$$

Mean Deviation

- The mean deviation of a set of values from a central value is the mean of absolute deviations of the values from the central value (Mean or Median or Mode).

Mean deviation about AM

For Raw data:

$$M.D.(\overline{X}) = \frac{\sum \left| x - \overline{x} \right|}{n}$$

For tabulated data:

$$M.D.(\overline{X}) = \frac{\sum f \left| x - \overline{x} \right|}{N}$$

Relative Measure:

$$Coeff\ of\ M.D.(\overline{X}) = \frac{M.D.(\overline{X})}{\overline{X}}$$

Mean deviation about Median

For Raw data:

$$M.D.(M) = \frac{\sum |x - M|}{n}$$

For tabulated data:

$$M.D.(M) = \frac{\sum f|x - M|}{N}$$

Relative Measure:

$$Coeff\ of\ M.D.(M) = \frac{M.D.(M)}{M}$$

Tip! (about Mean deviation)

Mean deviation can be calculated about any average - mean, mode, median, G.M., H.M., etc.

But, mean deviation is the least when it is measured from the median *(Minimal property of Median)*.

Therefore, **M.D.(M) should be preferred**.

Standard Deviation (S.D.)

- The standard deviation of a set of values is the positive square-root of mean of the squared deviations of the values from their Arithmetic Mean.

- It is denoted by **σ** (sigma).

S.D.: Formulae

	Raw data	Tabulated data
For Population	$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$	$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}}$
For Sample	$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{N-1}}$

Properties

1. S.D. is independent of the origin of measurement but not independent of scale.
2. Always, $\sigma \geq 0$.
3. S.D. is the least of all root-mean-square deviations.

Note: SD= σ ,Variance= σ^2

Coefficient of Variation (C.V.)

- Coefficient of Variation is relative measure of Standard deviation.

$$C.V. = \frac{Standard\ Deviation}{Arithmetic\ Mean} \times 100$$

$$C.V. = \frac{\sigma}{\bar{x}} \times 100$$

- It is independent of units of measurement. So, useful while comparing variation present in different frequency distributions.

Is Central Tendency and variation enough to describe frequency distributions?

Yes or No?

- **Clearly NO.**
- **There few measures need to be computed to describe data, such as;**
  - **Skewness**
  - **Kurtosis**

- **To compute skewness and kurtosis we need to know the “Moments”.**

Lets see!

Moments

- The characteristics of a frequency distributions are described by its Moments.
- Definition: The r^{th} moment of the set of values about any constant is the mean of r^{th} powers of the deviation of the values from the constant.
- Types of Moments:
  - Central Moments
  - Raw moments

Moments: Formulae

Here, *'a'* is any constant except AM.

	Central Moments	Raw Moments
Raw data	$\mu_r = \frac{\sum(x-\bar{x})^r}{n}$	$\mu'_r = \frac{\sum(x-a)^r}{n}$
Tabulated data	$\mu_r = \frac{\sum f(x-\bar{x})^r}{N}$	$\mu'_r = \frac{\sum f(x-a)^r}{N}$

Moments

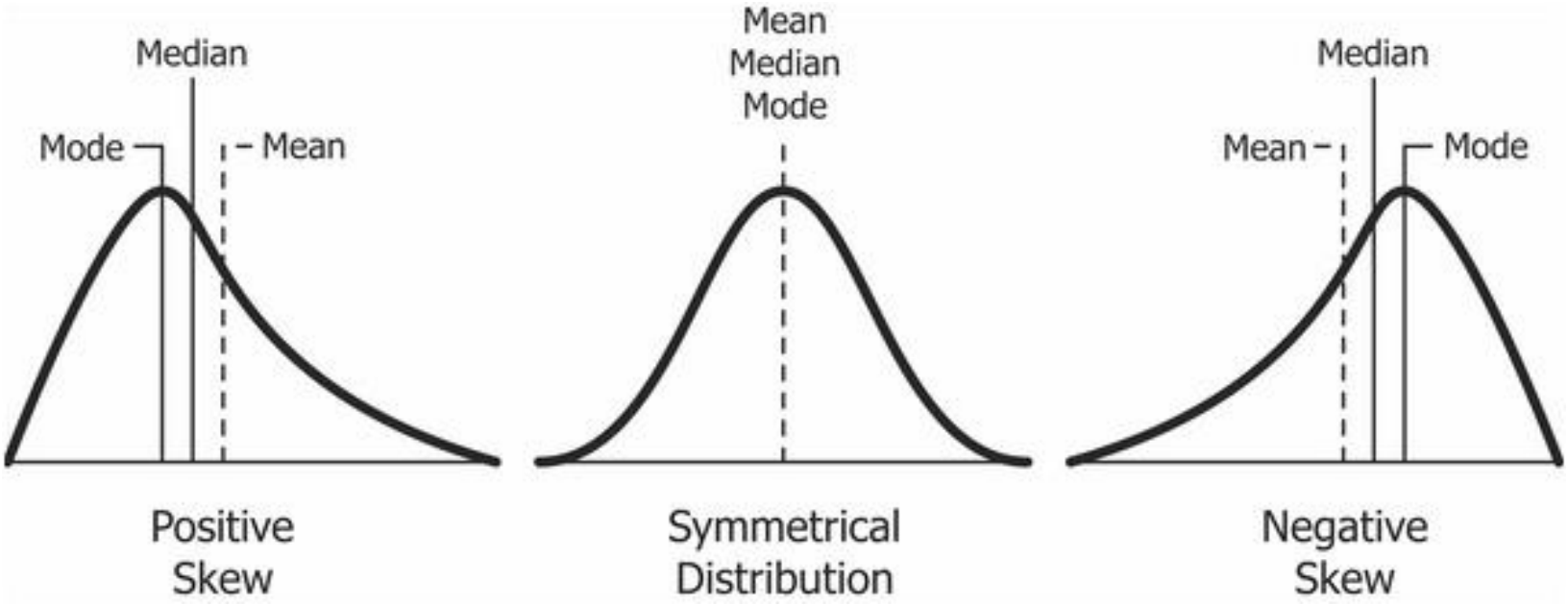
- Moments helps in measuring the scatteredness, asymmetry and peakedness of a curve for a particular distribution.

Applications of Moments

- The first moment about zero is the mean.
- The second moment about the mean is the sample variance.
- The third moment about the mean in calculating skewness.
- The fourth moment about the mean in the calculation of kurtosis.

Skewness

Skewness means non-symmetry or asymmetry or lack of symmetry.



Skewness

If *mean = median = mode*,

the shape of the distribution is **symmetric**.

If *mode < median < mean*,

the shape of the distribution trails to the right, is **positively skewed**.

If *mean < median < mode*,

the shape of the distribution trails to the left, is **negatively skewed.**

Measures of Skewness

1. Based on Moments
2. Karl Pearson's Coefficient of Skewness
3. Bowley's Coefficient of Skewness

1. Based on Moments

Formula:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{OR} \quad \gamma_1 = +\sqrt{\beta_1}$$

- Measure based on moments (i.e. β_1) is **more exact** in indicating degree of skewness of skewness than others.

2. Karl Pearson's Coefficient of Skewness

Formula:

$$S = \frac{mean - mode}{S.D.} = \frac{\bar{x} - Z}{\sigma}$$

Note:

If mode is ill-defined, we can go for following formula;

$$S = \frac{3(mean - median)}{S.D.} = \frac{3(\bar{x} - M)}{\sigma}$$

since we know, $Z = 3M - 2\bar{x}$

3. Bowley's Coefficient of Skewness

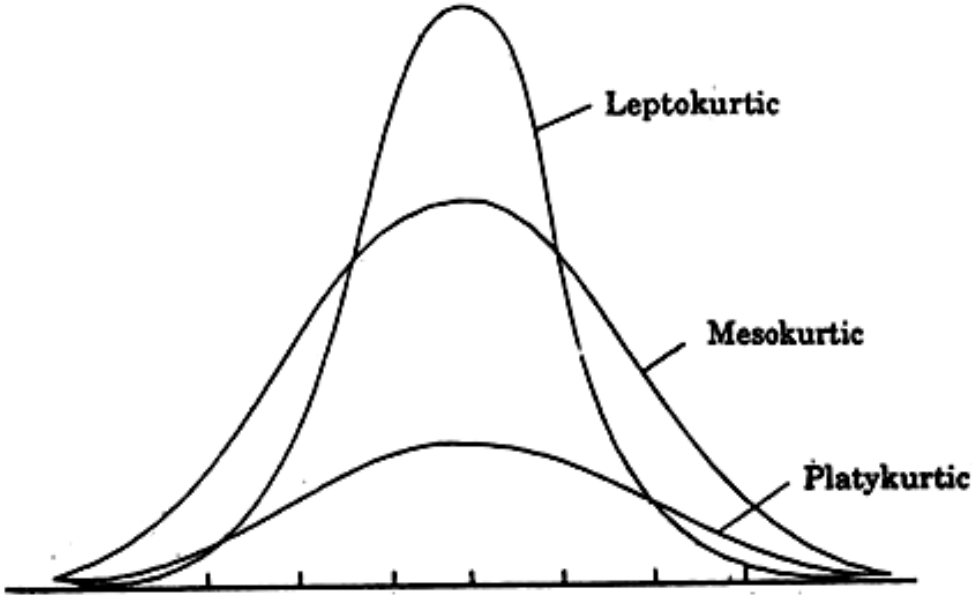
Bowley's Coefficient of Skewness is based on quartiles. Also called as *Galton skewness*.

Formula:

$$S = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

Kurtosis

Kurtosis is the degree of peakedness (non-flatness).



Kurtosis

Kurtosis: Based on Moments

Formula:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \quad \text{OR} \quad \gamma_1 = \beta_2 - 3$$

If,	Decision	If,
$\beta_2 = 3$	Platykurtic	$\gamma_1 = 0$
$\beta_2 < 3$	Mesokurtic	$\gamma_1 < 0$
$\beta_2 > 3$	Leptokurtic	$\gamma_1 > 0$

Skewness

- If skewness = 0, the data are perfectly symmetrical. But a skewness of exactly zero is quite unlikely for real-world data, so **how can you interpret the skewness number**?

Skewness

Bulmer, “Principles of Statistics” (1979) suggests the following rule of thumb:

- If skewness is less than −1 or greater than +1, the distribution is **highly skewed**.
- If skewness is between −1 and −½ or between +½ and +1, the distribution is **moderately skewed**.
- If skewness is between −½ and +½, the distribution is **approximately symmetric**.

Example:

Question: In frequency distribution, first four central moments are 0, 4, -2 and 2.4. Comment on Skewness and Kurtosis of the distribution.

Solution:

Given, $\mu_1 = 0$, $\mu_2 = 4$, $\mu_3 = -2$ and $\mu_4 = 2.4$

Therefore,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.0625 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = 0.15$$

Since, $\boldsymbol{\mu_3}$ **is negative**, the distribution is **negatively skewed**. Also, $\boldsymbol{\beta_1 = 0.0625}$ **is very small**, the distribution is **slightly skewed**.

Since, $\boldsymbol{\beta_1 = 0.15 < 3}$, the distribution is **platykurtic**.

Next Lecture

- Topic: **Probability**
  - **Introduction to Probability**
  - **Conditional probability**
  - **Bayes Theorem**
- Background material to study
  - Business Statistics, Chpt2, pp. 50
- We will have MCQ on this lecture: 10 Questions
- And will have recap of the lecture one.
- Discuss on the assignment of this lecture.

Thank You