Sane's STATS Academy of Statistics

# Preliminary Data Analysis

## Measures of Central Tendency

# We will be covering…

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
  - Quartiles
  - Options of central tendency in pandas
- Measures of Dispersion
  - Range
  - Semi Inter-Quartile Range
  - Mean Deviation
  - Variance
  - Standard Deviation
  - Coefficient of Variation
  - Skewness
  - Kurtosis
  - Options of dispersion in pandas

# What are averages?

- These are statistical constants which enable us to comprehend in a single effort the significance of the whole thing.

# Measures of Central Tendency

- Mean
  - Arithmetic mean
- Median
- Mode
- Quartiles , Deciles and Percentiles

# Arithmetic mean

- Arithmetic mean of a given set of observations is their sum divided by the number of observations.
- e.g. A.M. of 5, 8, 10, 15, 24 and 28 is

$$\frac{5+8+10+15+24+28}{6} = \frac{90}{6} = 15$$

# Median

- Median is that value which divides the set of given numbers in two equal parts.
- E.g. Median of numbers 5, 10, 8, 15, 28 and 24 can be calculated as follows:
  - Arrange the given numbers in ascending/descending order as 5, 8, 10, 15, 24, 28
  - Count the numbers. They are 6 in number.
  - The middle two numbers are 10 and 15. Hence median is the arithmetic mean of 10 and 15. i.e. 12.5

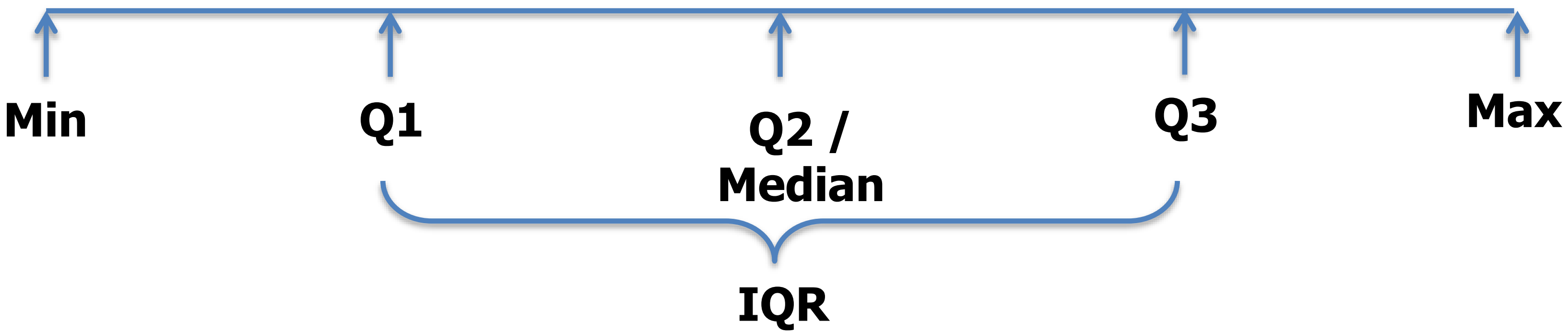
# Mode

- The mode is the value which has greatest frequency.
- E.g. Mode of numbers [ 4, 5, 5, 6, 7, 8, 6 , 5 ] is 5

# Quartiles

- Quartiles divide the given data into four equal parts.

**Data represented as a line**



- Inter-quartile range (IQR) is given by the formula:

$$\boxed{IQR = Q3 - Q1}$$

Sane's STATS Academy of Statistics

# Preliminary Data Analysis

Measures of Dispersion

# Measures of Dispersion

- Absolute Measures
  - Range
  - Quartile Deviation or Semi-Interquartile Range
  - Mean Deviation
  - Standard Deviation
- Relative Measures
  - Coefficient of Variation

# Range

- Range is defined as the difference between the two extreme observations in a distribution (i.e. greatest (maximum) and the smallest (minimum) observation.)
- E.g. Range of 5, 8, 10, 15, 24 and 28 is 28 – 5 = 23

# Quartile Deviation or Semi-Interquartile Range

- Quartile Deviation: It is calculated by a formula:

$$QD = \frac{Q_3 - Q_1}{2}$$

# Mean Deviation

- Mean Deviation (average deviation) is a measure of dispersion that is obtained on taking the average (arithmetic mean) of the absolute deviation of the given values from a measure of central tendency (mean).

# Standard Deviation

- Standard Deviation is defined as the positive square root of the arithmetic mean of the squares of the deviations of the given observations from their arithmetic mean.
- More is the magnitude of a standard deviation more is the dispersion.
- e.g. Data with SD=23.4 can be said to be more dispersed than data with  SD=12.7.

# Coefficient of Variation

- The ratio of SD and Mean
- CV is unit less quantity

$$CV = \frac{SD}{Mean} * 100$$

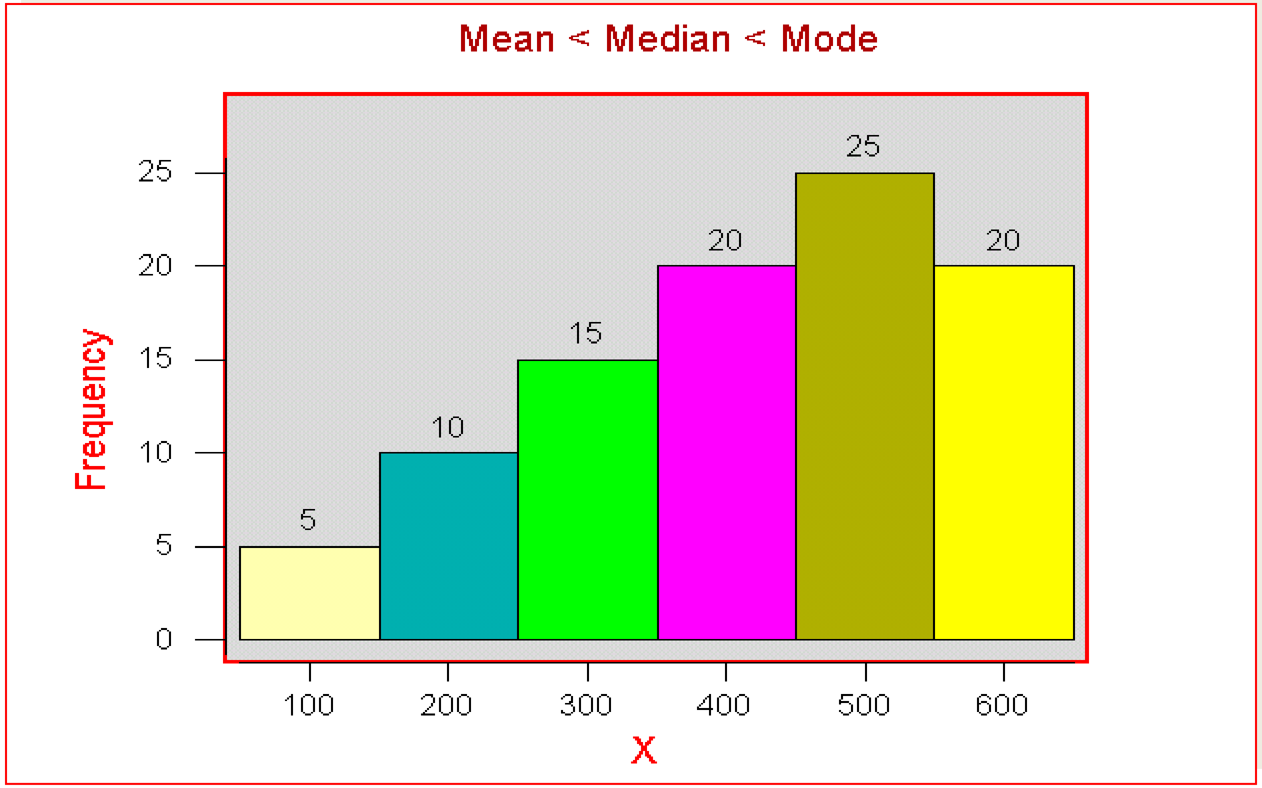
$$CV = \frac{\sigma}{\mu} * 100$$

# Skewness and Kurtosis

- **Skewness**
  - Measure of asymmetry of a frequency distribution
    - Skewed to left
    - Symmetric or unskewed
    - Skewed to right
- **Kurtosis**
  - Measure of flatness or peakedness of a frequency distribution
    - Platykurtic (relatively flat)
    - Mesokurtic (normal)
    - Leptokurtic (relatively peaked)

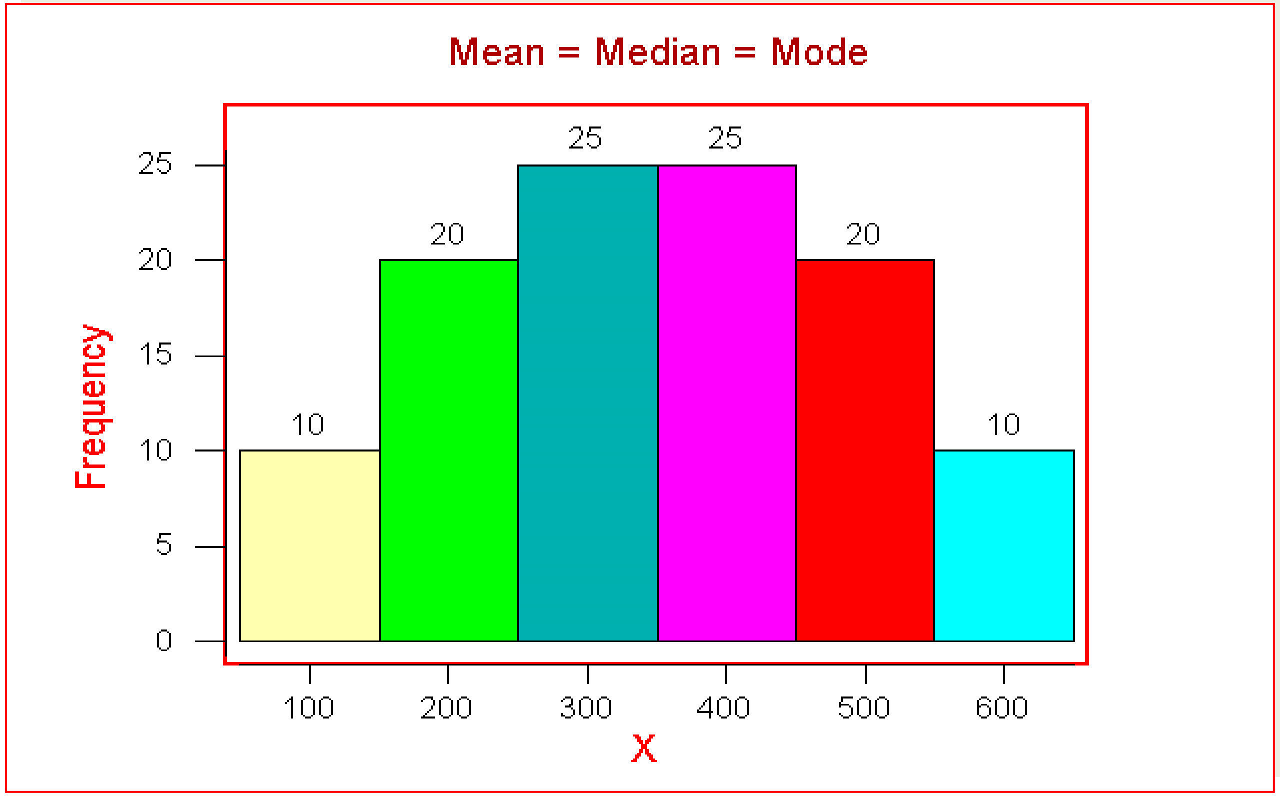
# Skewness

## Skewed to left



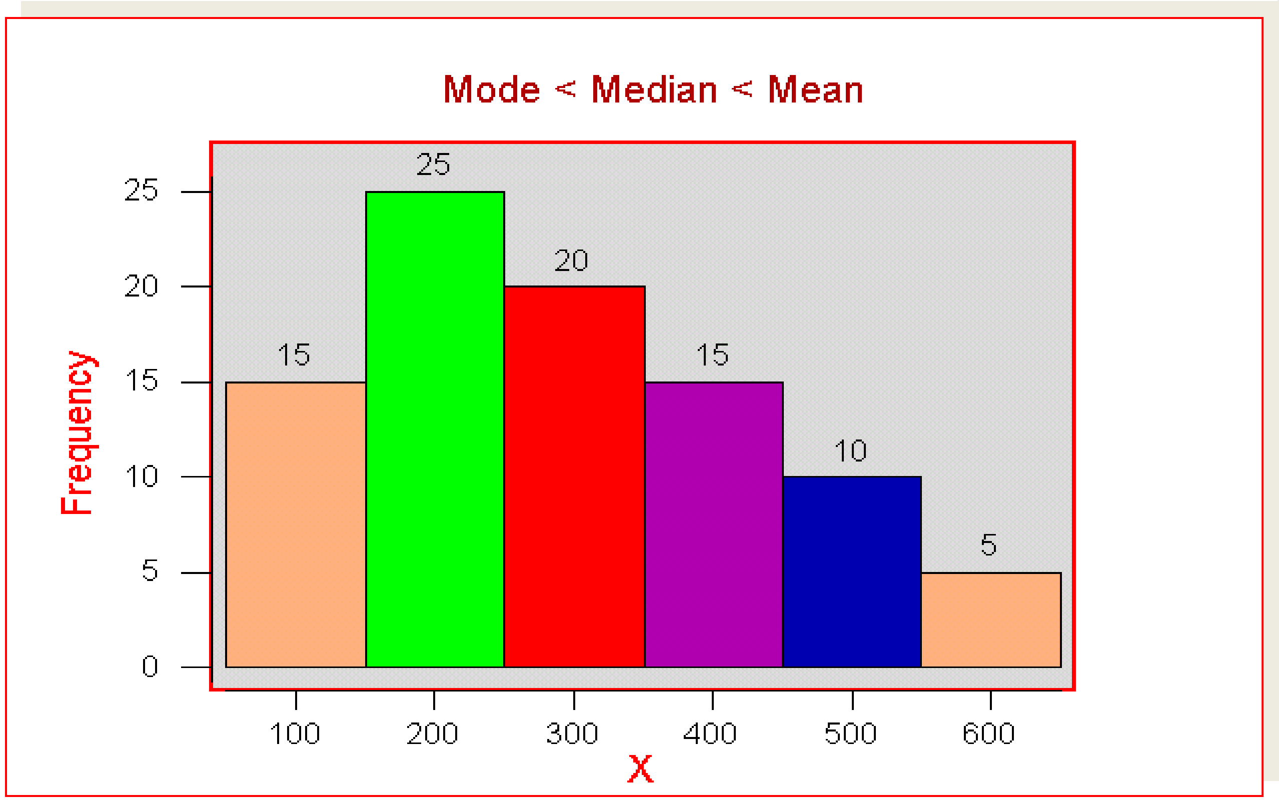
# Skewness

## Symmetric



# Skewness

## Skewed to right



# Coefficient of Skewness

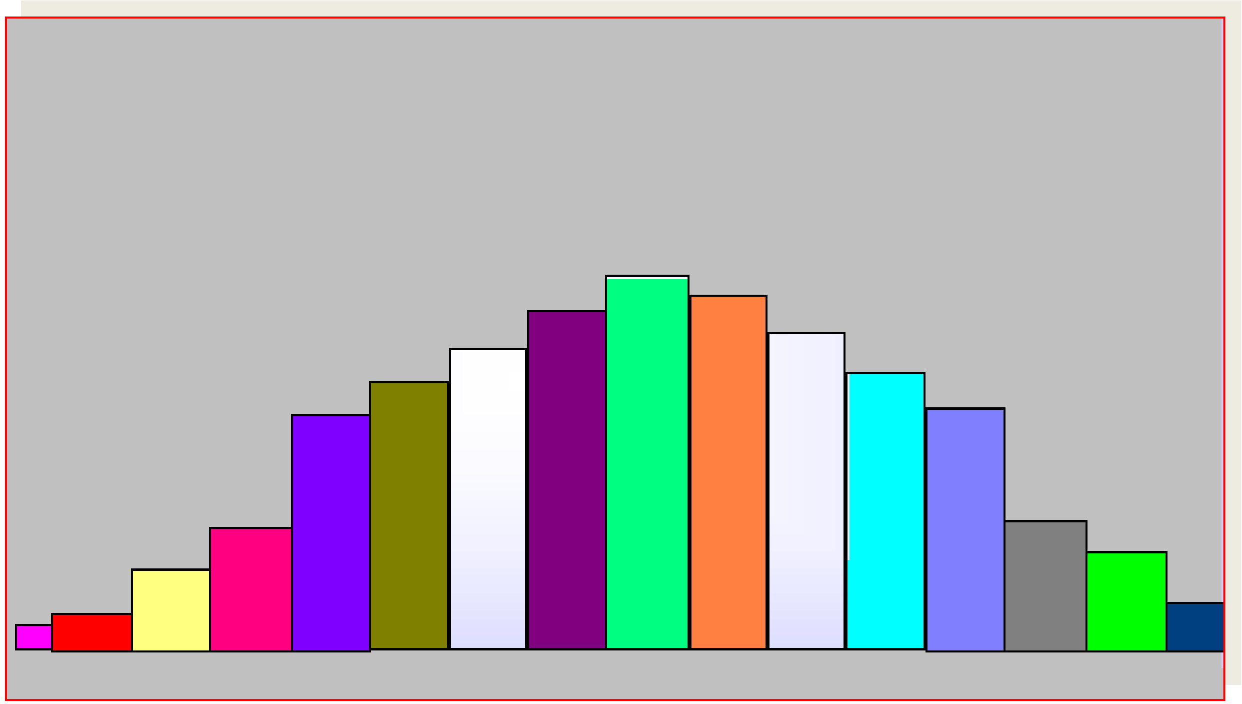
- Coefficient of Skewness (CS):

$$\boxed{CS = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i \;\; - \mu)^3}{\sigma^3}}$$

  - CS is negative for left-skewed data.
  - CS is positive for right-skewed data.
  - $|CS| > 1$ suggests high degree of skewness.
  - $0.5 \le |CS| \le 1$ suggests moderate skewness.
  - $|CS| < 0.5$ suggests relative symmetry.

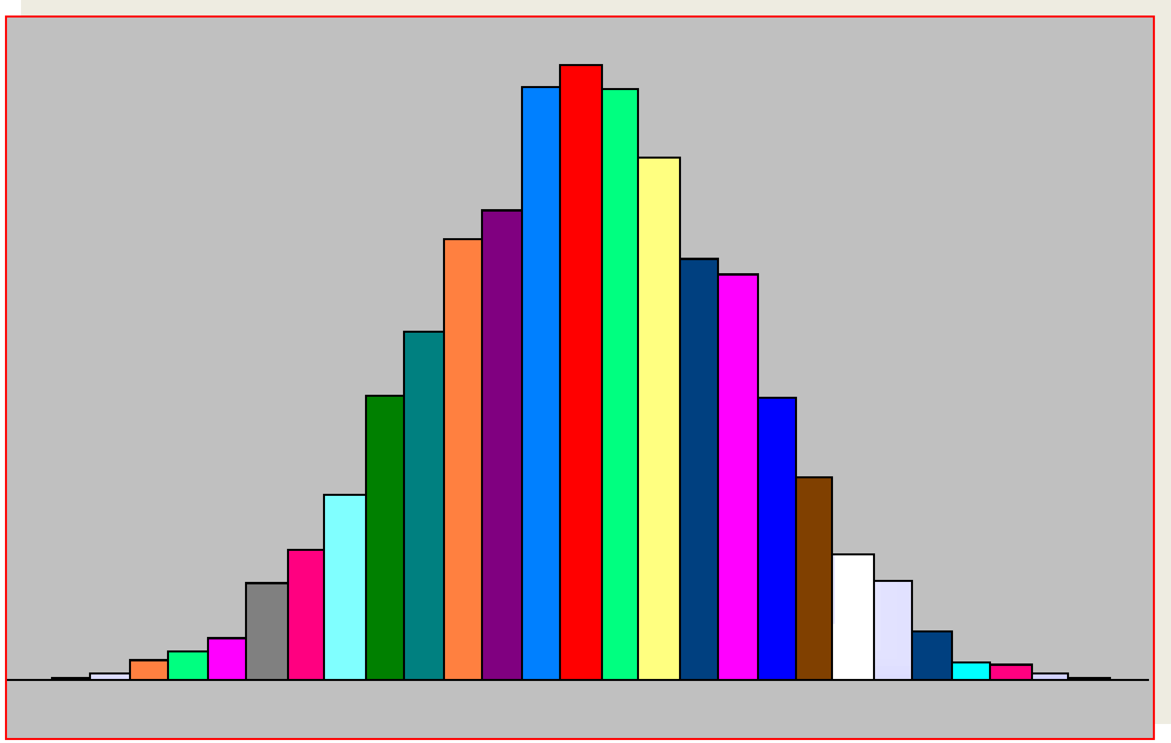
# Kurtosis

**Platykurtic** - flat distribution



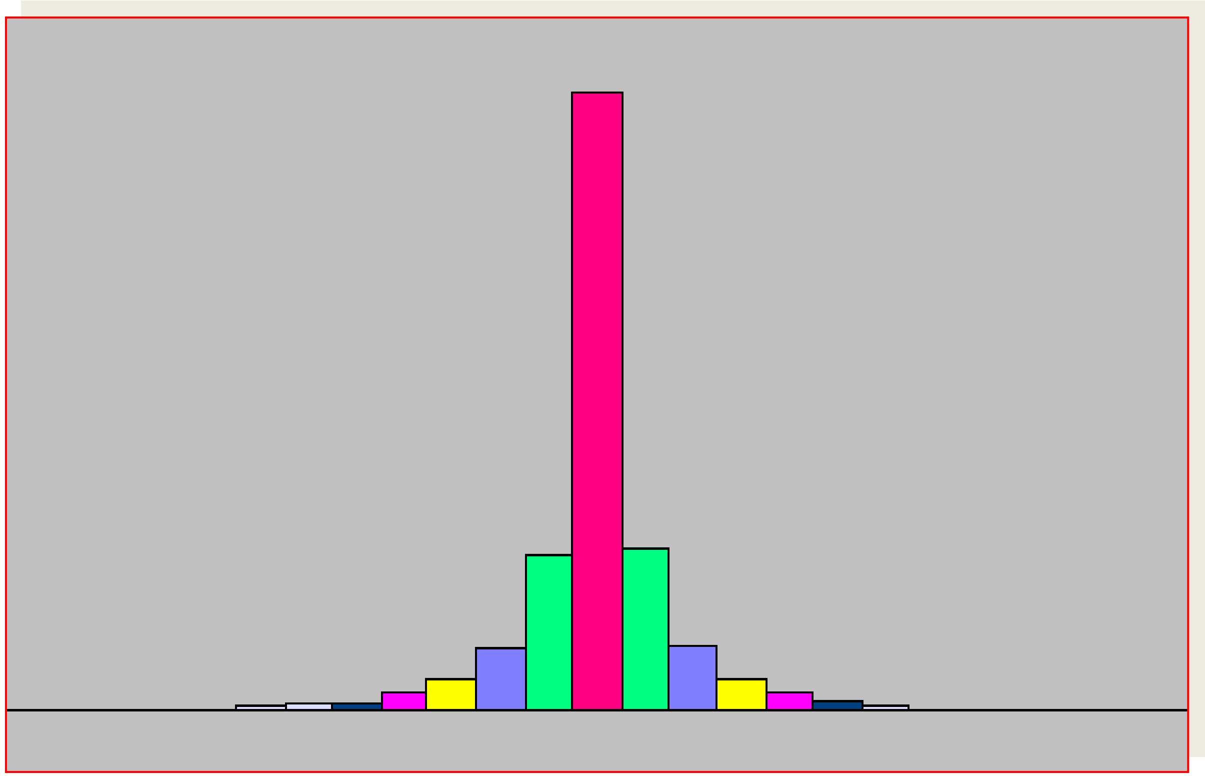
# Kurtosis

**Mesokurtic** - not too flat and not too peaked



# Kurtosis

**Leptokurtic** - peaked distribution



# Kurtosis

- **Kurtosis** refers to the peakedness (i.e., high, narrow) or flatness (i.e., short, flat-topped) of a histogram.
- The coefficient of kurtosis (CK) measures the degree of kurtosis of a population

$$\boxed{CK = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i \; -\mu)^4}{\sigma^4} - 3}$$

- CK < 0 indicates the data is somewhat flat with a wide degree of dispersion.
- CK > 0 indicates the data is somewhat peaked with less dispersion.