SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

III Year / VI Semester

Unit 2 – DESCRIPTIVE ANALYTICS USING STATISTICS

Topic 5 : Percentiles and Moments

Percentile

- In a dataset, a percentile is the point at which x% of the values are less than the value at that point.

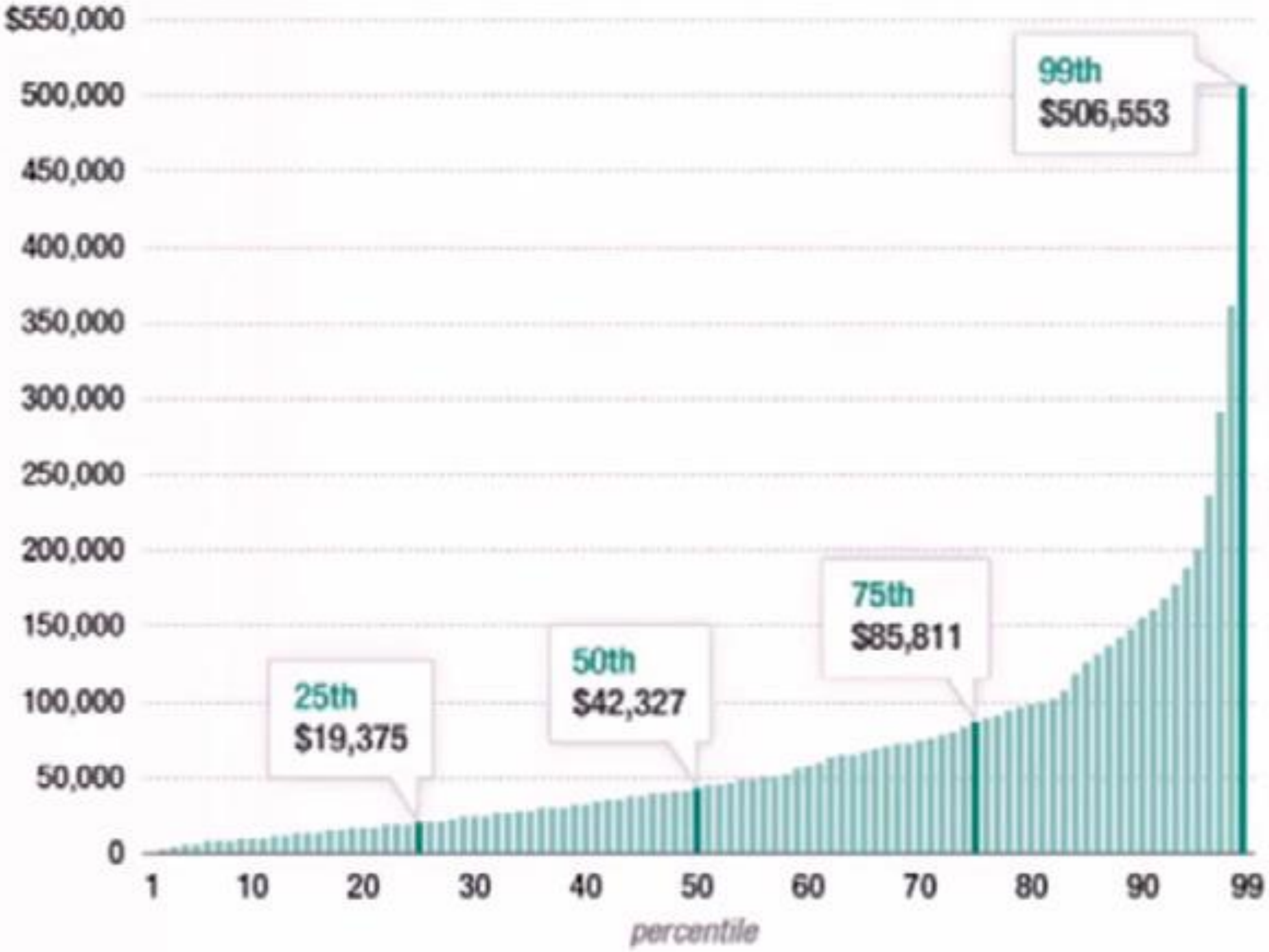
- A percentile is a measure at which that percentage of the total values are the same as or below that Measure.

- Percentiles are useful for giving the relative standing of particular data in a dataset. Percentiles are essentially normalized ranks.

Example

- The 80th percentile is a value where you'll find 80% of the values lower and 20% of the values higher.

Percentile


$550,000
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
1
10
20
30
40
50
60
70
80
90
99
percentile
25th
$19,375
50th
$42,327
75th
$85,811
99th
$506,553

Percentile

Quartiles

- Quartiles divide the data into four groups, each containing an equal number of values.

- Quartiles are divided by the 25th, 50th, and 75th percentile, also called the first, second and third quartile. It can be represented as Q1,Q2,Q3 and Q4 respectively.

- One quarter of the values are less than or equal to the 25th percentile.

- Three quarters of the values are less than or equal to the 75th percentile.

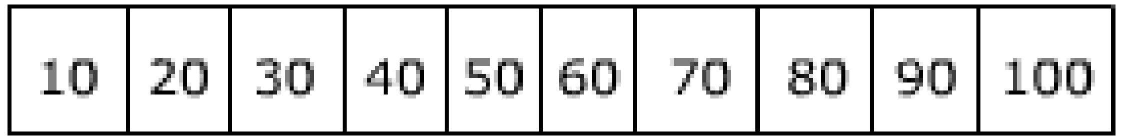
Interquartile range

- The difference between the 75th (Q3) and 25th (Q1) percentile is called the interquartile range.

- For example, the interquartile range (IQR), when we talk about a distribution, is the area in the middle of the distribution that contains 50% of the values.

Percentile

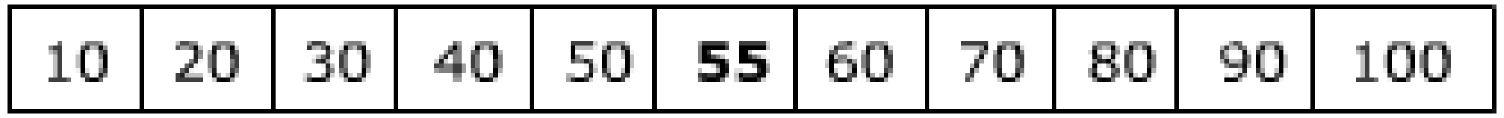
Example



10	20	30	40	50	60	70	80	90	100

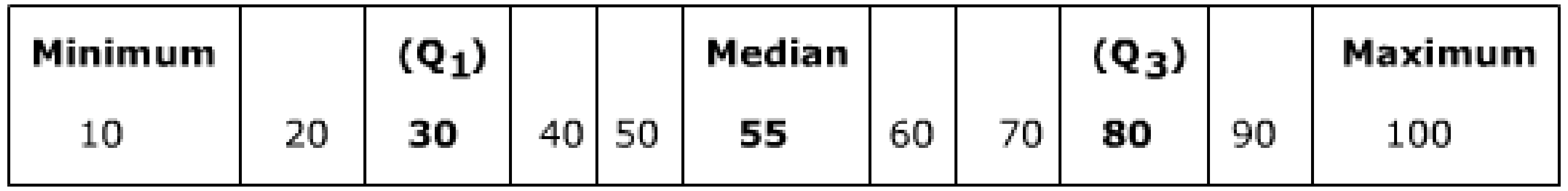
In the above dataset the minimum value is 13 and maximum value is 100.

Median is 55.



10	20	30	40	50	**55**	60	70	80	90	100

- The first quartile (Q1) is just the "median" of all the values to the left of the true median.
- We can see that 30 is the middle number of the numbers to the left of the true median, so 30 is the 25th percentile and the first quartile (Q1).
- What if we were asked for the 75th percentile? We know that the 75th percentile is the third quartile (Q3). The third quartile (Q3) is similarly the "median" of the values to the right of the true median.
- We can see that 80 is the middle number of the numbers to the right of the true median, so 80 is the 75th percentile and the third quartile (Q3).



Minimum		**(Q_1)**			**Median**			**(Q_3)**		**Maximum**
10	20	**30**	40	50	**55**	60	70	**80**	90	100

Moments

- Moments can be defined as quantitative measures of the shape of a probability density function.

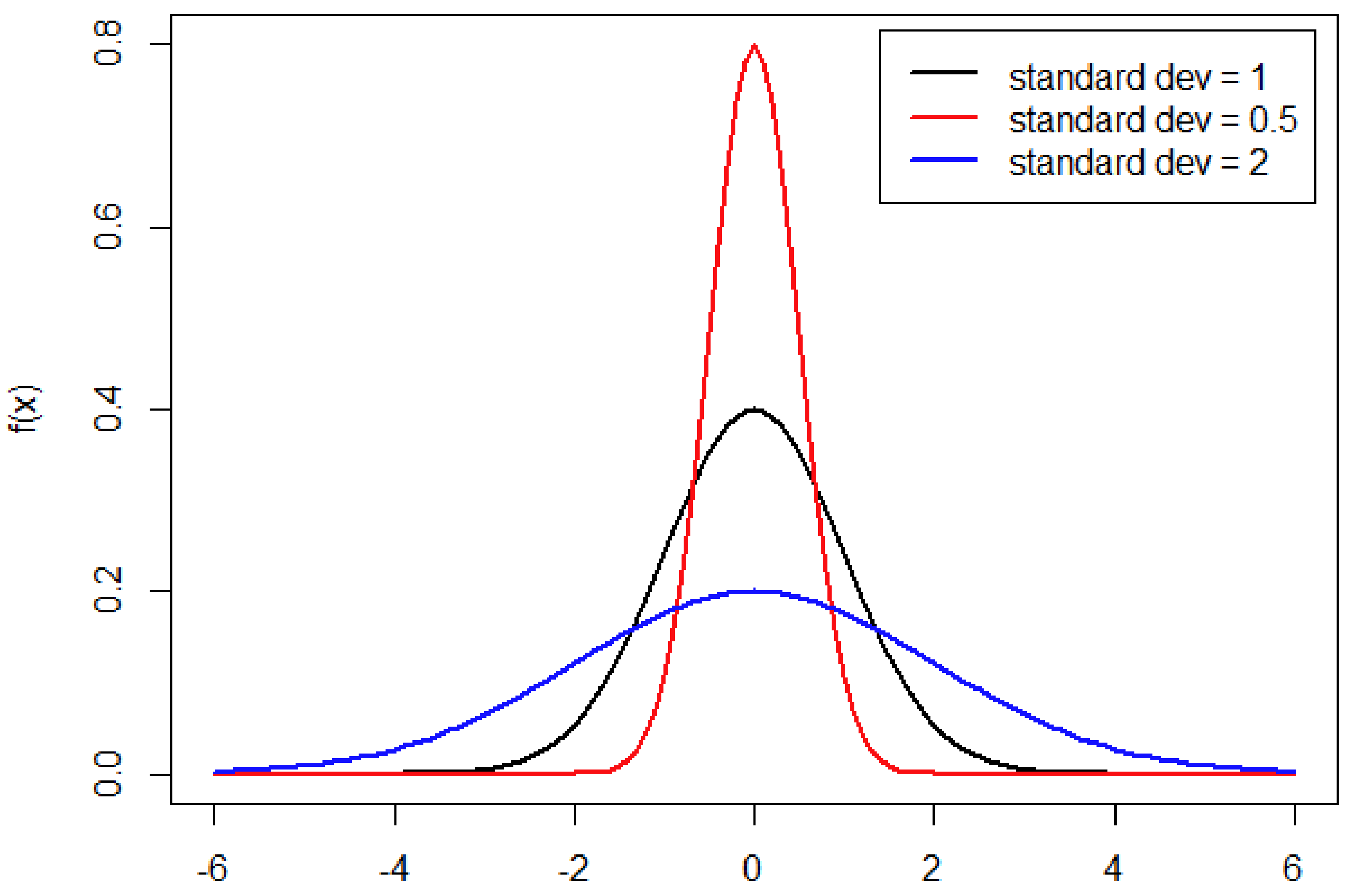
- Moments in statistics are popularly used to describe the characteristic of a distribution. The shape of any distribution can be described by its various ‘moments’.

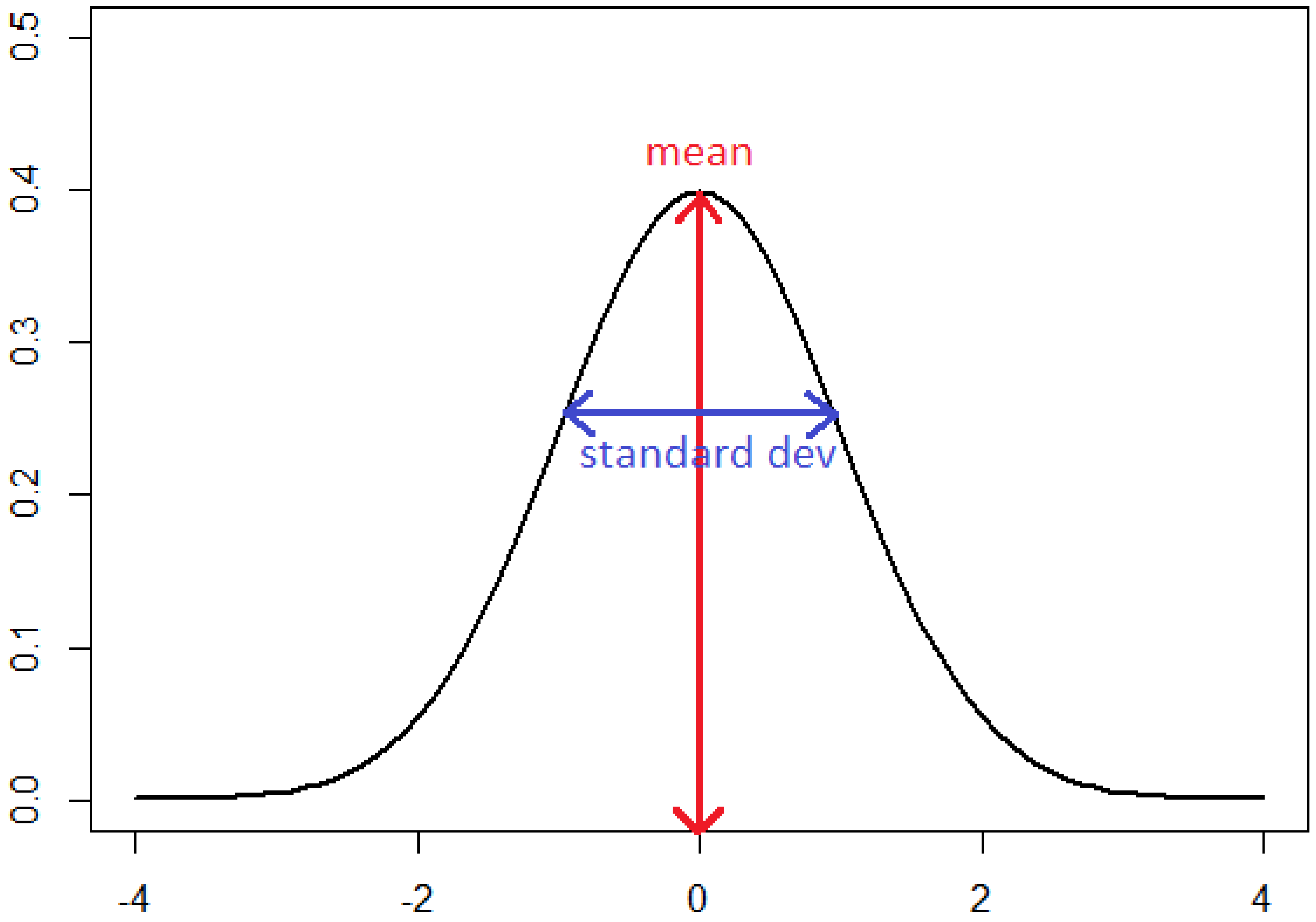
The four commonly used moments in statistics are-

- The first moment - the mean (Measure the location of the central point)
- The second moment - variance (which indicates the width or deviation)
- The third moment – skewness (which indicates any asymmetric ‘leaning’ to either left or right)
- The fouth moment - kurtosis (which indicates the degree of central ‘peakedness’ or, equivalently, the ‘fatness’ of the outer tails.)

Moments

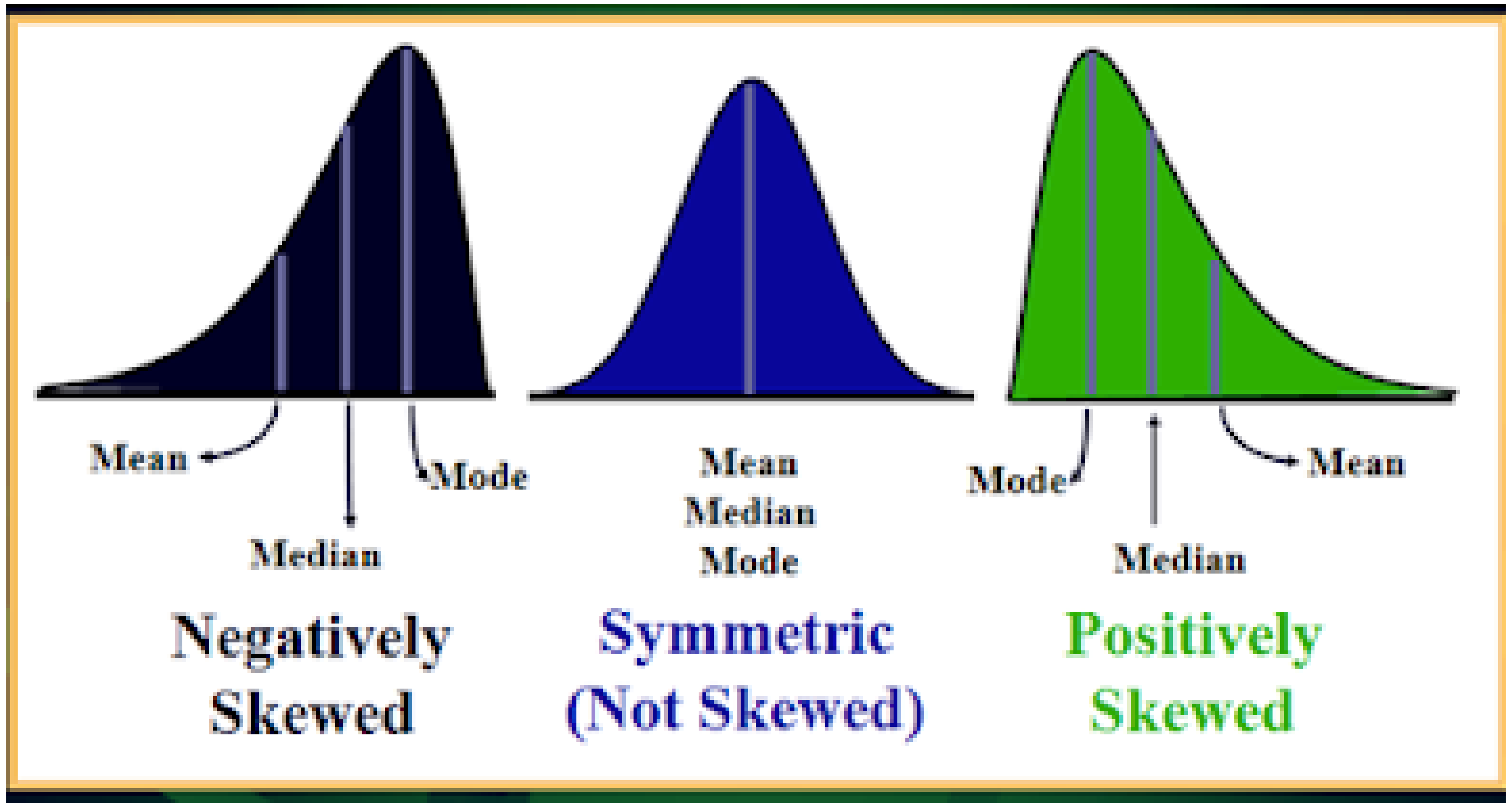
The greater the variance/ standard deviation (e.g. blue line), the wider the spread of values around the mean. If a variance is lower, the values are cumulated closer to the mean (red line) and the peak is higher.





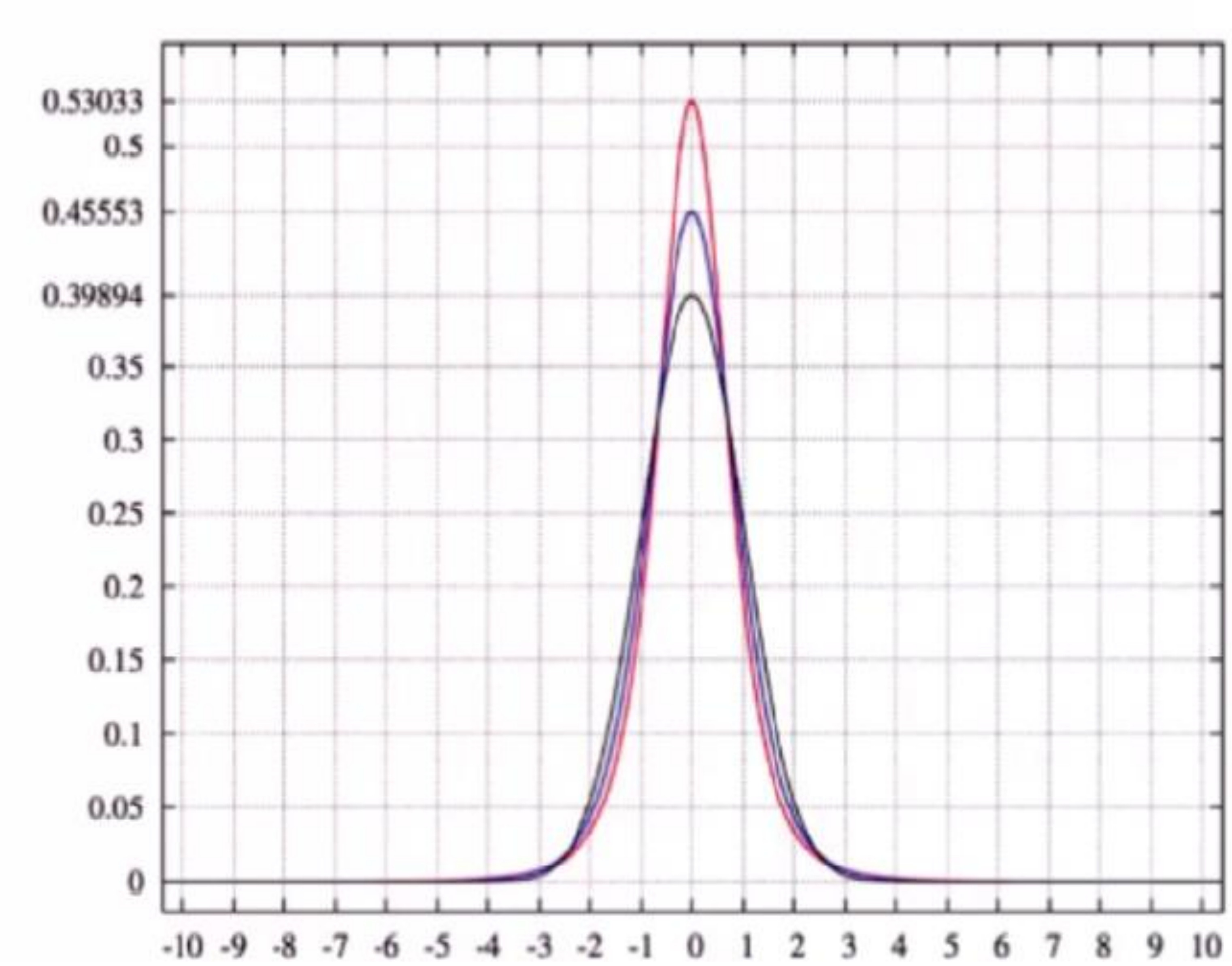
Moments

- Skew is how lopsided the data is, how stretched out one of the tails might be.
- Symmetrical distribution: as in examples above. Both tails are symmetrical and the skewness is equal to zero.
- Positive skew (right-skewed, right-tailed, skewed to the right): the right tail (with larger values) is longer.
- Negative skewed (left-skewed, left-tailed, skewed to the left): the left tail (with small values) is longer.



Moments

- Kurtosis is how peaked, how squished together the data distribution is.
- It focuses on the tails of the distribution and explains whether the distribution is flat or rather with a high peak. Kurtosis informs us whether our distribution is richer in extreme values than normal distribution.



Covariance

- In statistics, covariance is the measure of the directional relationship between two random variables.

- These are ways of measuring whether two different attributes are related to each other in a set of data, which can be a very useful thing to find out.

- A positive covariance indicates that both random variables tend to move upward or downward at the same time.

- A negative covariance indicates that both variables tend to move away from each other — when one moves upward the other moves downward, and vice versa.

- Covariance between 2 random variables is calculated by taking the product of the difference between the value of each random variable and its mean, summing all the products, and finally dividing it by the number of values in the dataset.

$$cov_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation

Correlation

- The correlation between two random variables measures both the strength and direction of a linear relationship that exists between them.
- The Pearson Correlation Coefficient is defined to be the covariance of x and y divided by the product of each random variable’s standard deviation.

$$r = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}$$

$$\frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1} \frac{\sum (y_i - \bar{y})^2}{n-1}}}$$

Correlation

Correlation

- Correlation of -1 means there's a perfect inverse correlation, so as one value increases, the other decreases, and vice versa.

- A correlation of 0 means there's no correlation at all between these two sets of attributes.

- A correlation of 1 would imply perfect correlation, where these two attributes are moving in exactly the same way as you look at different data points.

THANK YOU