Descriptive statistics

Dr. P. Kalpana, M.E., PhD.
Faculty of Mechanical Engineering
IIITDM Kancheepuram

Objectives

- ➢Statistics
- ➢Types of statistics
- ➢Measures of Central tendency
- ➢Measures of Variability.
- ➢ Measures of Divergence from Normality.
- ➢Measures of Probability

What is statistics

- The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions
- Statistics is a science that helps us make better decisions in business and economics as well as in other fields
- Statistics teaches us how to summarize, analyze, and draw meaningful inferences from data that then lead to improve decisions
- These decisions that we make help us improve the running, for example, a department, a company, the entire economy, etc

Categories of statistics

Descriptive Statistics

✓Collect
✓Organize
✓Analyze
✓Summarize
✓Display

Inferential Statistics

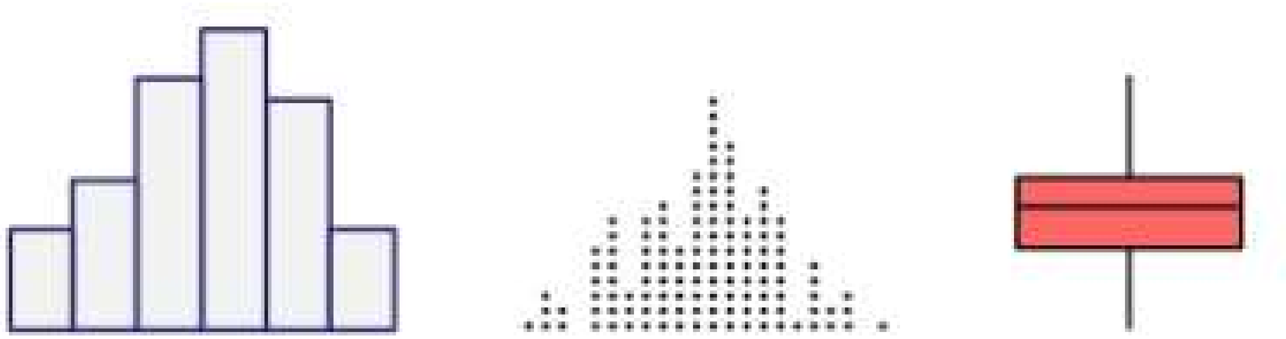
✓Predict and forecast value of population parameters
✓Test hypothesis about value of population parameter based on sample statistic
✓Make decision

Descriptive Statistics

- Methods of describing, summarizing and organizing of a data set
- Allow us to make the sense of data
- Helps exploring and making conclusions about the data in order to make rational decisions
- Includes calculating things such as the average of the data, its spread and the shape it produces
- Example :
  - The weight of a product in a production line
  - The time taken to process an application
- Graphical Displays:
  - Often used along with the quantitative measures to enable clarity of communication
  - When analyzing a graphical display, conclusions can be drawn based on several characteristics of the graph

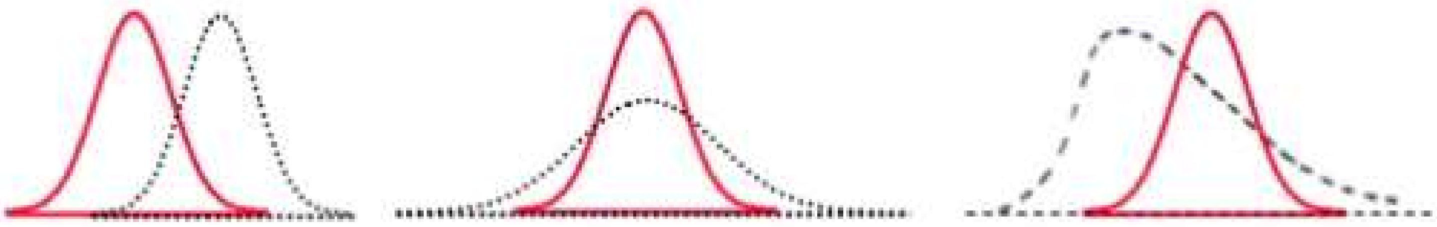
Descriptive Statistics

- You may ask questions such as
    - Where is the approximate middle or center of the graph ?
    - How spread out are the data values on the graph?
    - What is the overall shape of the graph?
    - Does it have any interesting patterns?



Descriptive Statistics

- The following measures are used to describe a data set
- Measures of position (also referred to as central tendency or location measures)
- Measures of spread (also referred to as variability or dispersion measures)
- Measures of shapes



Descriptive Statistics

Measures of central tendency

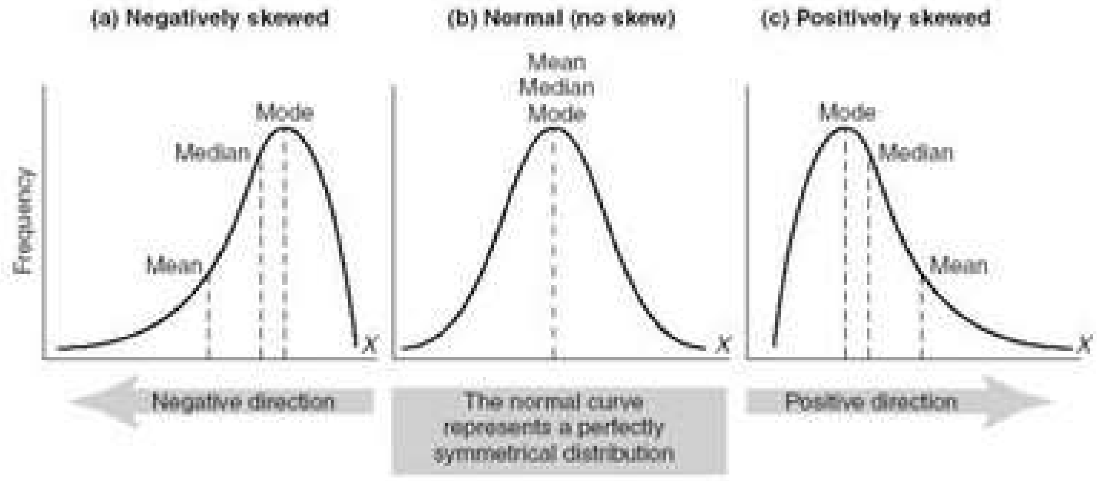
- ✓Arithmetic mean
- ✓Weighted mean
- ✓Geometric mean
- ✓Median
- ✓Mode
- ✓Percentiles and Quartiles

Measures of Dispersion

- ✓Skewness
- ✓ Kurtosis
- ✓ Range
- ✓ Inter quartile range
- ✓Variance
- ✓Standard deviation
- ✓Coefficient of variation

Measures Of Central Tendency And Location

- Yield information about the centre or middle part of a group of numbers
- It refers to where the data is centered
- A single number to describe the data
- Measures of central tendency are usually computed from sample data rather than from population data
- Major types of measures
  - Mean
  - Median
  - Mode

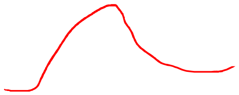


Examples of normal and skewed distributions

Arithmetic Mean

- The **arithmetic mean (or simply *mean*) *of a set of data is the sum of the data*** values divided by the number of observations.
- It is otherwise called as average
- It is commonly used as statistic position

$$\boxed{\bar{x}} = \frac{5+10+3+8+6}{5}$$

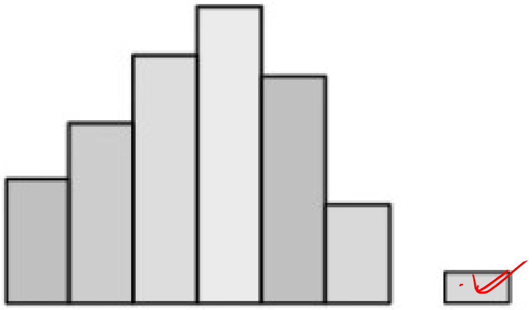


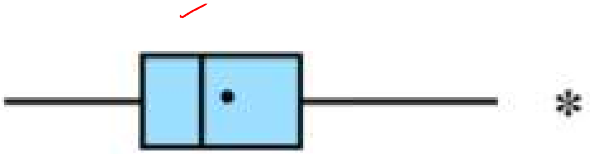
- Applicable to interval and ratio data
- Not applicable to nominal and ordinal data
- Affected each value in the data set including extreme data points
- It works well when the distribution is symmetric and there are no outliers
- The mean of a sample is denoted by x-bar while the mean of a population is denoted by "µ" (parameter)

$\bar{x}$ μ

Outliers

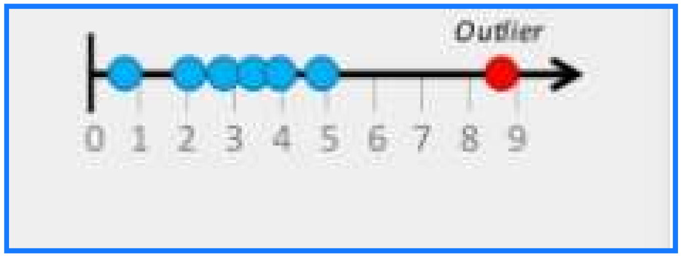
- A data point significantly greater or smaller than other data points in the data set
- It is useful when analyzing data to identify outliers
- They may affect the calculation of descriptive statistics
- Outliers can occur In any given data set and in any distribution
- The easiest way to detect them is by graphing the data or graphical method such as
  - Histograms
  - Box plots
  - Normal probability plots
- Outliers may indicate an experimental error or incorrect recording of data





Outliers

- They may also occur by chance
- It may be normal to have high or low data points
- You need to decide whether to exclude them before carrying out your analysis
- An outlier should be excluded if it due to measurement or human error
- This example is about the time taken to process a sample of applications



- It is clear that one data point is far distant from the rest of the values. This point is an outlier

Arithmetic Mean

Population mean

$$\boxed{\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}}$$

where N = population size

Sample mean

$$\boxed{\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

where n = sample size. The mean is appropriate for numerical data.

Example: Mean for ungrouped Data

- The data show the number of patients in a sample of six hospitals who acquired infection while hospitalized . Find the mean

110 76 29 38 105 31

Solution

$$\bar{X} = \sum X = \frac{110 + 76 + 29 + 38 + 105 + 31}{6} = 64.8$$

Example: Mean for grouped Data

- Example: You grew fifty baby carrots using special soil. You dig them up and measure their lengths (to the nearest mm) and group the results:

Length (mm)	Frequency
150 - 154	5
155 - 159	2
160 - 164	6
165 - 169	8
170 - 174	9
175 - 179	11
180 - 184	6
185 - 189	3

Mean for grouped Data

Length (mm)	Midpoint x	Frequency f	fx
150 - 154	152	5	760
155 - 159	157	2	314
160 - 164	162	6	972
165 - 169	167	8	1336
170 - 174	172	9	1548
175 - 179	177	11	1947
180 - 184	182	6	1092
185 - 189	187	3	561
	Totals:	50	8530

Estimated Mean $\mu = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$

Estimated Mean $= \frac{8530}{50} =$ **170.6 mm**

Median

- The middle value where exactly half of the data values are above it and half are below it.
- It is less widely used but a useful statistic due to its robustness.
- it can reduce the effect of outliers and often used when the data is nonsymmetrical.
- It is important to ensure that the values are ordered before calculating the median.
- Remember also that with an even number of values, the median is the mean of the two middle values.
- It can be computed for ratio, interval and ordinal data
- The median will be the number located in the $0.50(n+1)$th ordered position.

The ages of a sample of five college students are given
21 25 19 20 22
Arranging the data in the ascending order give
19 20 21 22 25

Thus the median is 21

The ages of a sample of four data points are given
76 73 80 85
Arranging the data in the ascending order give
73 76 80 85

Thus the median is 78

Example: Median for grouped Data

For the median of grouped data, we find the cumulative frequencies and then calculate the median number n/2. The median lies in the group (class) which corresponds to the cumulative frequency in which n/2 lies.

$$Median = l + \frac{h}{f}\left(\frac{n}{2} - c\right)$$

- Here

l= Lower class boundary of the ~~model~~ median class
f= Frequency of the median class
n=∑f= Number of ~~values~~ observation or total frequency
c= Cumulative frequency of the class preceding the median class
h= Class interval size of the model class

The Median is the mean of the 25th and the 26th length

Example: Median for grouped Data

Group	f	Class Boundary	Cumulative Frequency
149.5 150 - 154 .5	5 ✓	149.5-154.5	5
155 - 159	2	154.5-159.5	7
160 - 164	6	159.5-164.5	13
165 - 169	8	164.5-169.5	21 ✓
170 - 174 ✓	9 ✓	169.5-174.5	30 → median class
175 - 179	11	174.5-179.5	41
180 - 184	6	179.5-184.5	47
185 - 189	3	184.5-189	50

l= ~~169.5~~ 169.5

f=9

n=50

c= 21

h= 5

$$\text{Estimated Median}=169.5+\frac{\left(\frac{50}{2}\right)-21}{9}\times 5 \qquad =169.5+2.22=171.72$$

50

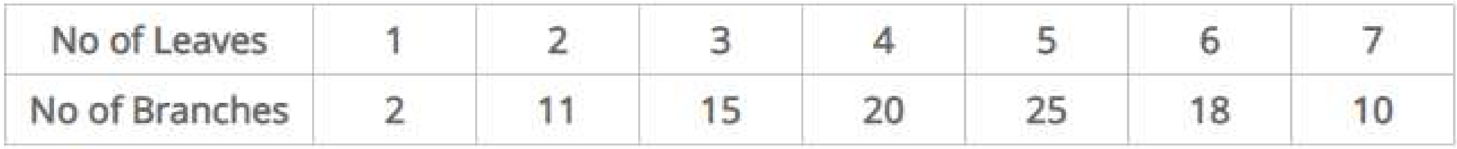
25, 26

position of median = ~~50~~ 0.5(50+1)

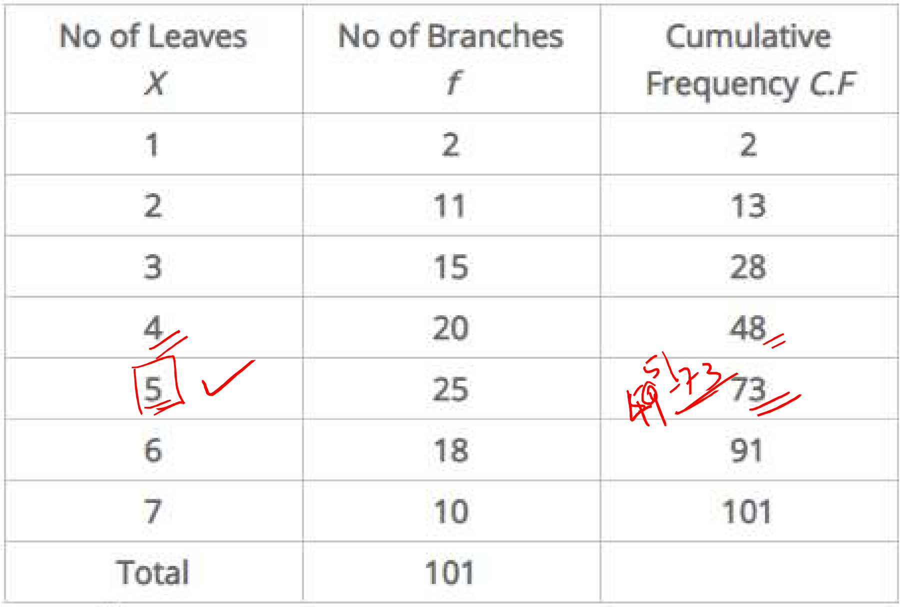
25.5

Median from Discrete Data

- The following frequency distribution is classified according to the number of leaves on different branches. Calculate the median number of leaves per branch.



No of Leaves	1	2	3	4	5	6	7
No of Branches	2	11	15	20	25	18	10



No of Leaves *X*	No of Branches *f*	Cumulative Frequency *C.F*
1	2	2
2	11	13
3	15	28
4	20	48
5 ✓	25	49 - 51 - 73 73
6	18	91
7	10	101
Total	101	

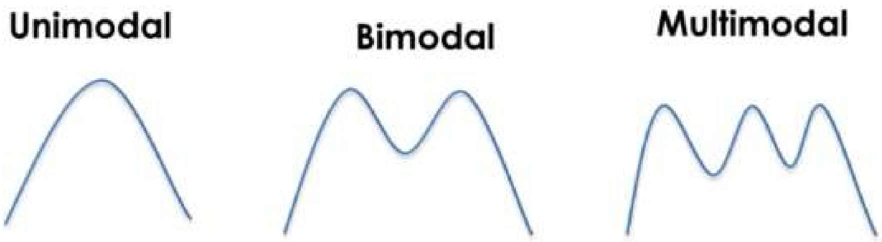
$$\text{Median} = \text{Size of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \frac{101+1}{2} = \frac{102}{2} = 51 \text{ item}$$

$$\text{Median} = \mathbf{5} \text{ because } 51^{th} \text{ item corresponds to } \mathbf{5}$$

Mode - Ungrouped data

- The **Mode** is the value that occurs the most often in a data set
- It is rarely used as a central tendency measure
- It is more useful however to distinguish between unimodal and multimodal distributions when data has more than one peak
- The mode is most commonly used with categorical data (Nominal)
- Example:

Number of television sets	Frequency
1	13
2	18
3	1
4	10
5	2



Data point which occurs most frequently is 2 as it has a frequency of 18. So the **mode** is 2

Mode of grouped data

With frequency distribution with equal class interval sizes, the class which has the maximum frequency is called the model class.

$$Mode = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

Take the same example

Here
The Modal group is the one with the highest frequency, which is **175 - 179**
l= Lower class boundary of the model class=174.5
fm= Frequency of the model class (maximum frequency)=11
f1= Frequency preceding the model class frequency=9
f2= Frequency following the model class frequency=6
h= Class interval size of the model class=5

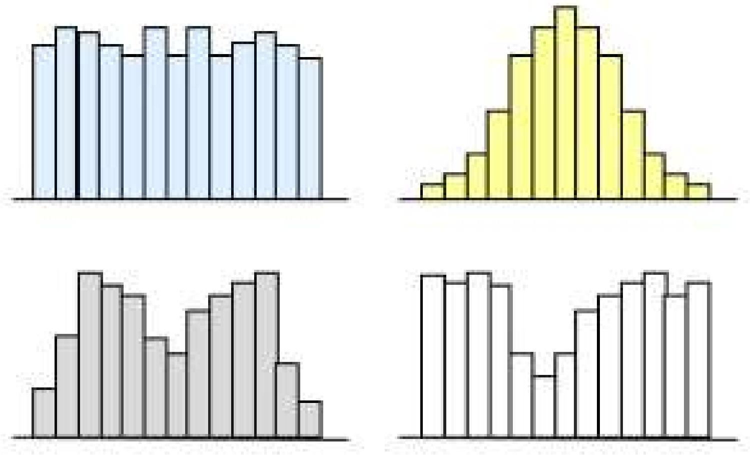
Estimated Mode=175.9

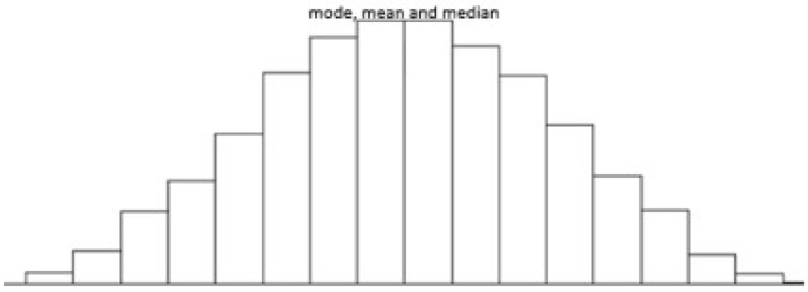
Shape of distribution

- The shape can reveal a lot of information about the data
- The shape can also help identifying which descriptive statistic is more appropriate to use in a given situation.
- If the data is symmetrical for example, then it is appropriate to use the mean or median to measure the central tendency as they are almost equal.
- If the data is skewed, then the median will be a more appropriate to measure the central tendency.
- The two common statistics that measure the shape of the data are
    - Skewness
    - Kurtosis.

Shape of distribution-Symmetric

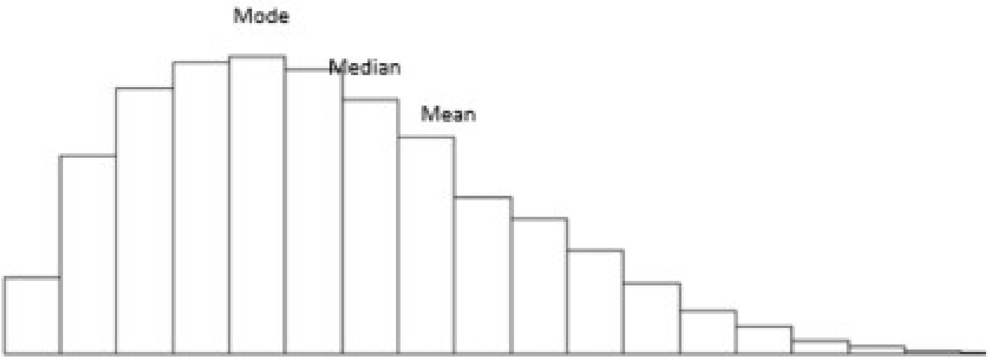
- if the center of the data divides a graph of the distribution into two mirror images,
- so that the portion on one side of the middle is nearly identical to the portion on the other side, the distribution is said to be symmetric.
- Skewness is zero
- Ex: uniform, normal (bell shaped), camel-back, and bow-tie shaped





Shape of distribution-Asymmetric or skewed

Skewness describes whether the data points are distributed symmetrically around the mean

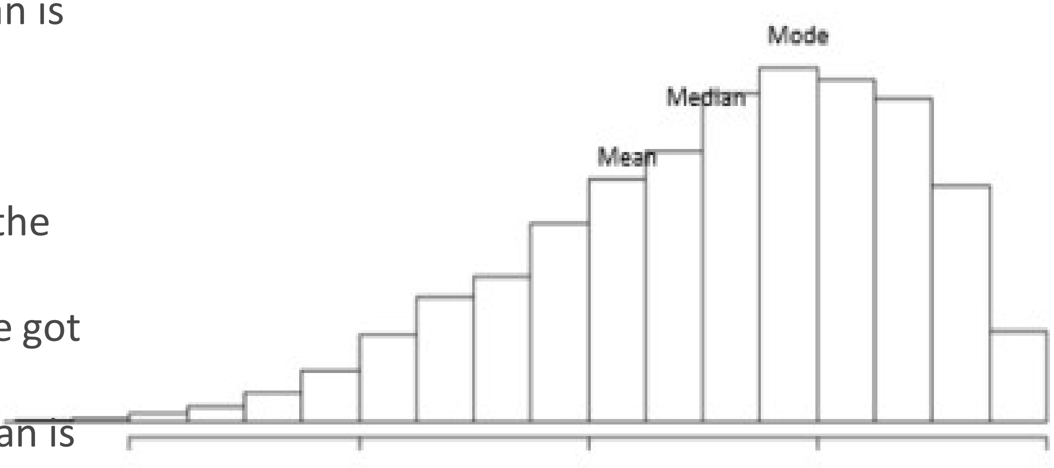


Positive or right skewed:

➢Skewness is positive is a distribution is skewed to the right

➢Eg: Tough question paper less number of people got highest marks

➢For continuous numerical unimodal data the mean is usually greater than the median



Negtively or left skewed:

➢Skewness is negative for distributions skewed to the left

➢Eg: Easy question Paper – more number of people got highest marks

➢For continuous numerical unimodal data, the mean is usually less than the median

If it is skewed data you go for median as the center tendency

Coefficient of skewness

- Measure of skewness (Karl pearson coeeficient)

$$S = \frac{3(\mu - M_d)}{\sigma}$$

- S skewness coefficient
- μ mean of the dataset
- M_d median of the data set
- σ standard deviation of the data set

- If S<0, the distribution is negatively skewed (skew to the left)
- If S=0, the distribution is symmetric (not skewed)
- If S>0, the distribution is positively skewed

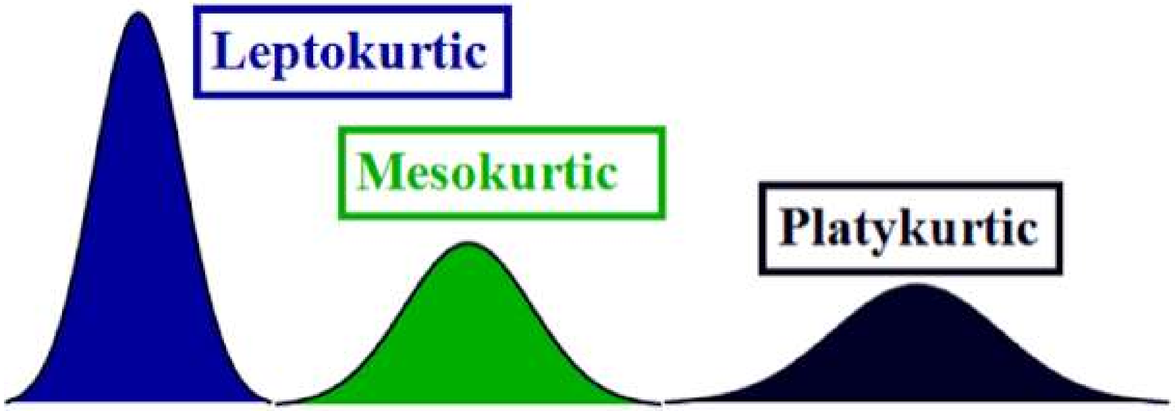
$\mu = 23; Md = 26;$ σ=12.3
S=-0.73

$\mu = 26; Md = 26;$ σ=12.3
S=0

$\mu = 29; Md = 26;$ σ=12.3
S=0.73

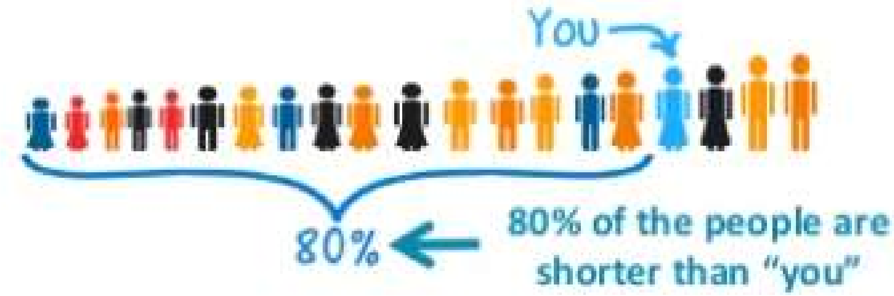
Shape of distribution-Kurtosis

- Measures the degree of flatness (or peakness of shape)
- When the data values are clustered around the middle then the distribution is more peaked.
  - A greater kurtosis value (Leptokurtic)
- When the data values are spread around more evenly , then the distribution is more flatted
  - A smaller kurtosis value (platykurtic)
- Normal curve is neither very peaked nor very flat topped. So it is taken as the basis for comparison
  - The normal curve is called Mesokurtic



Measures of location, or position- Percentiles

- Percentiles and quartiles are measures that indicate the location or position of a value relative to the entire set of data.
- Mainly used in educational and health related fields to indicate the position of an individual in a group
- It divides the group of data into 100 parts
- Example: 80th percentile indicates that the most 80%data lie below it , and at least 20% data lie above it



- The median and 50th percentile have the same value
- Applicable for ordinal, interval and ratio data
- Not applicable for nominal data
- Kth percentile = value located in the (K/100)*(n + 1)th ordered position

Procedure for calculating percentile

- Procedure 1:
- Step1 : Arrange the Numbers in ascending order
- Step 2: Find out the location for the given kth percentile

$$i = \frac{k}{100}(n+1)$$

- *i* is the location corresponding to the kth percentile
- If *i* is whole number , the number in that location will be the kth percentile value
- If the number is not a whole number we can split the number as integer (I) and Decimal D
- The percentile value=Value in the Ith position+ D*(I+1 th postion value-Ith Postion Value)

Percentiles- Example

Procedure1:

Raw data:

14,12,19,23,5, 13,28,17

Ordered arrays:

5,12,13, 14,17,19,23,28

Find out the Location of 30th Percentile and 50th percentile?

Location of 30th Percentile= (30/100)(8+1)=2.7

The percentile value=Value in the Ith position+ D*(I+1 th postion value-Ith Postion Value)

I=2 and D=0.7

The value of 30th Percentile is = 12+0.7(13-12)=12.7

Location of 50^{h} Percentile= (50/100)(8+1)=4.5

The value of 50th Percentile is 14+0.5(17-13)=15.5

Procedure for calculating percentile

Procedure 2:

Step 1: arrange data in numerical order from low to high

Step 2: Calculated the index i=(p/100)*n

p the given percentile

n number of data in the data set

- If i is not a whole number, the answer is in the next position of i
- i=2.1 the location of the correspond to the percentile p is 3
- i=3.9 the location of the correspond to the percentile p is 4
- If i is a whole number, the answer is the average of numbers in i^{th} position and $i+1^{th}$ position

- In the previous Example for 30 percentile location index i=2.7
- So the next whole number is 3.
- The 30 percentile is at 3rd position in the array .
- Value at 30percentile is 13.

Percentile Rank

It indicates the percent of all scores that fall below a particular score

The Percentile Formula for the selected Number in the data set is given as

$$\boxed{Percentile\ Rank = \frac{(B + 0.5E)}{n} \times 100}$$

B Number of scores below a given score

E Number of scores equal to the given score, including the given score. If there are no other scores equal to the given score then E=0

n total number of scores

The higher the percentile rank , the better the score when compared to the others scores

The lower the percentile rank , the poorer the score when compared to the others scores

Percentile Rank: Example

- The following is a set of 32 marks achieved by students on an examination worth 100 marks

18	42	52	59	68	73	83	89
27	45	53	61	70	75	83	90
38	45	56	64	72	82	85	90
40	48	58	67	72	83	85	97

- Determine the percentile ranking of a score 72

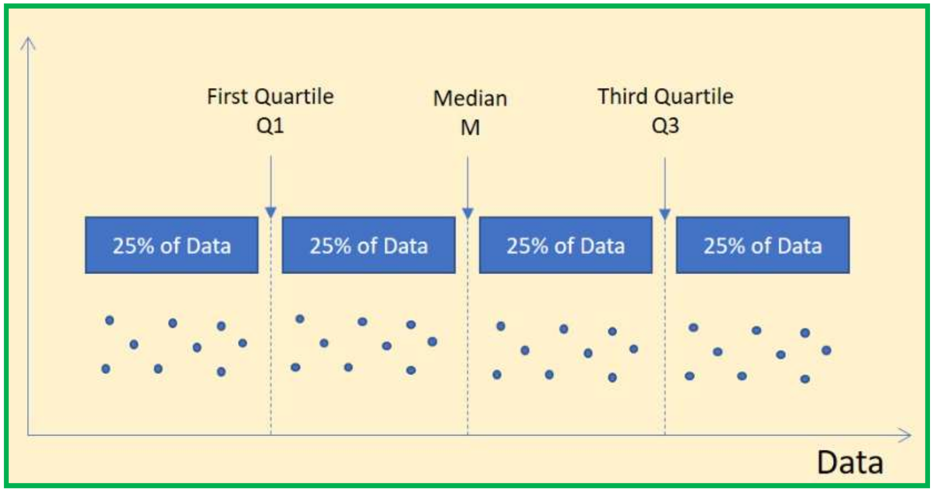
$$Percentile\ Rank = \frac{(B + 0.5E)}{n} \times 100$$

$$= \frac{18+(0.5\times 2)}{32} \times 100 = 59.4 \approx 60\ percentile$$

Quartiles

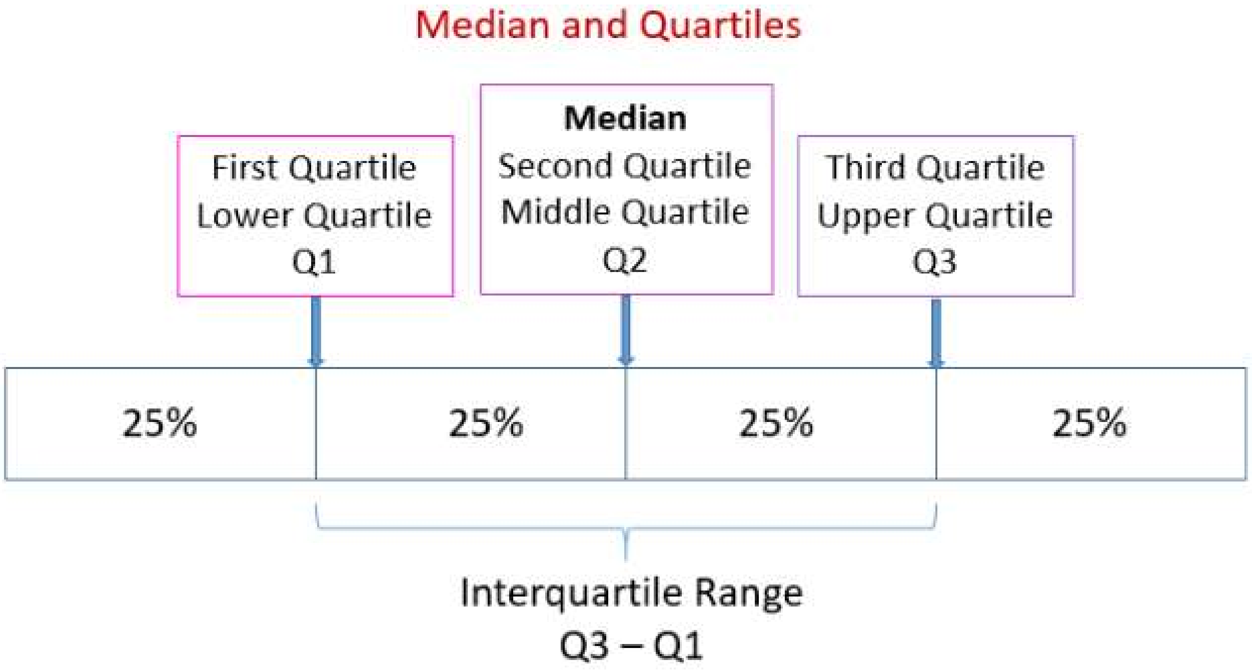
- **Quartiles are descriptive measures that separate large data sets into four** quarters.
- The **first quartile, *Q1, (or 25th percentile) separates approximately the smallest 25%*** of the data from the remainder of the data.
- The **second quartile** *Q2, (or 50th percentile) is the median*
- The **third quartile, *Q3, (or 75th percentile), separates approximately the smallest 75%*** of the data from the remaining largest 25% of the data.

$$\begin{aligned} Q_1 &= \text{the value in the } \mathbf{0.25}(n+1)\text{th ordered position} \\ Q_2 &= \text{the value in the } \mathbf{0.50}(n+1)\text{th ordered position} \\ Q_3 &= \text{the value in the } \mathbf{0.75}(n+1)\text{th ordered position} \end{aligned}$$



Inter Quartile Range

- The term Inter quartile Range (IQR) refers to the difference between Q3 and Q1 (IQR = Q3 – Q1).
- IQR statistic is more robust with respect to outliers
- It removes the outliers



Quartiles

- Find the median, lower quartile and upper quartile of the following numbers.

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

Solution:

First, arrange the data in ascending order:

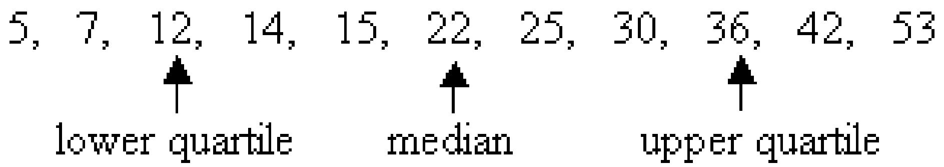
5 7 12 14 15 22 25 30 36 42 53

Find out the location index of Q1,Q2 and Q3

Location index for Q1=0.25×(11+1)=3

Location index for Q2=0.5×(11+1)=6

Location index for Q3=0.75×(11+1)=9



- Median (middle value) = 22
- Lower quartile (middle value of the lower half) = 12
- Upper quartile (middle value of the upper half) = 36

Quartiles- Example

If there is an even number of data items, then we need to get the average of the middle numbers.

Find the median, lower quartile, upper quartile, inter quartile range and range of the following numbers.

12 5 22 30 7 36 14 42 15 53 25 65

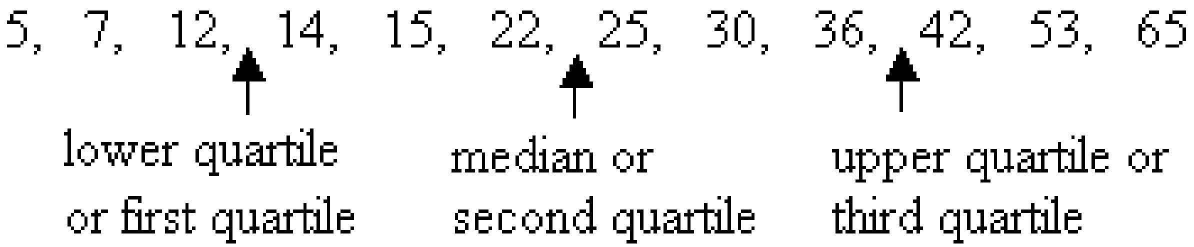
First, arrange the data in ascending order

Find out the location index of Q1,Q2 and Q3

Location index for Q1=0.25×(12+1)=3.25

Location index for Q2=0.5×(12+1)=6.5

Location index for Q3=0.75×(12+1)=9.75



5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53, 65

lower quartile or first quartile

median or second quartile

upper quartile or third quartile

Quartiles- Example

- Lower quartile or first quartile = (12+14)/2=13
- Median or second quartile = (22+25)/2=23.5
- Upper quartile or third quartile = (36+42)/2=39
- Inter quartile range = Upper quartile – lower quartile= 39 – 13 = 26
- Range = largest value – smallest value = 65 – 5 = 60

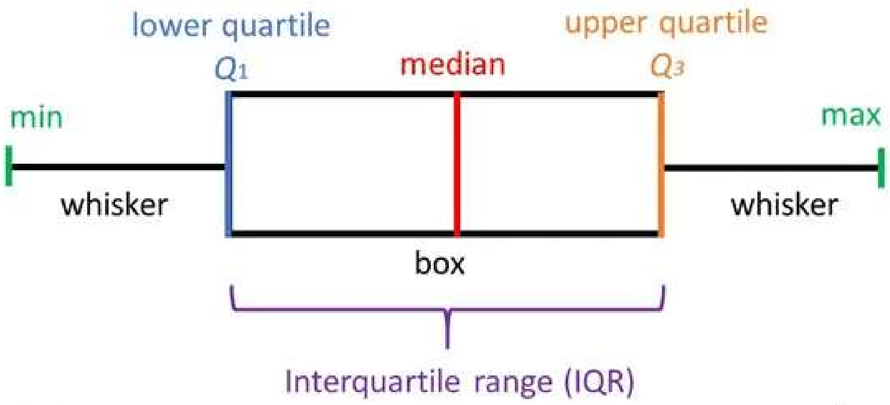
- Five-Number Summary
- The five-number summary refers to the five descriptive measures: minimum, first quartile, median, third quartile, and maximum.
- We can present a graph of the five-number summary called a box-and-whisker plot

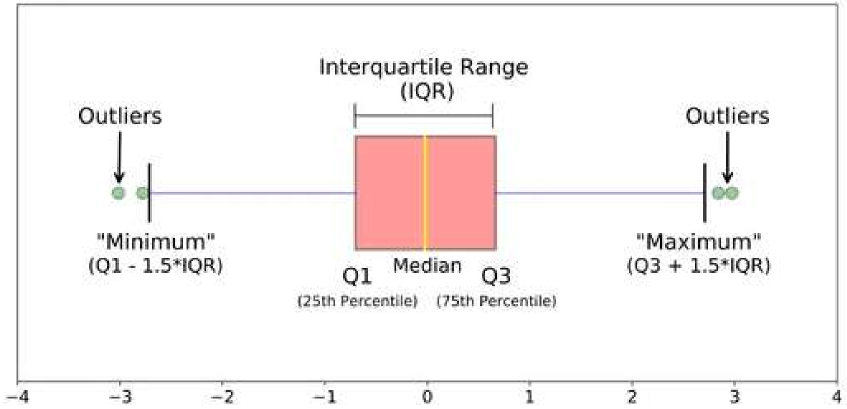
$$\boxed{\text{minimum} < Q_1 < \text{median} < Q_3 < \text{maximum}}$$

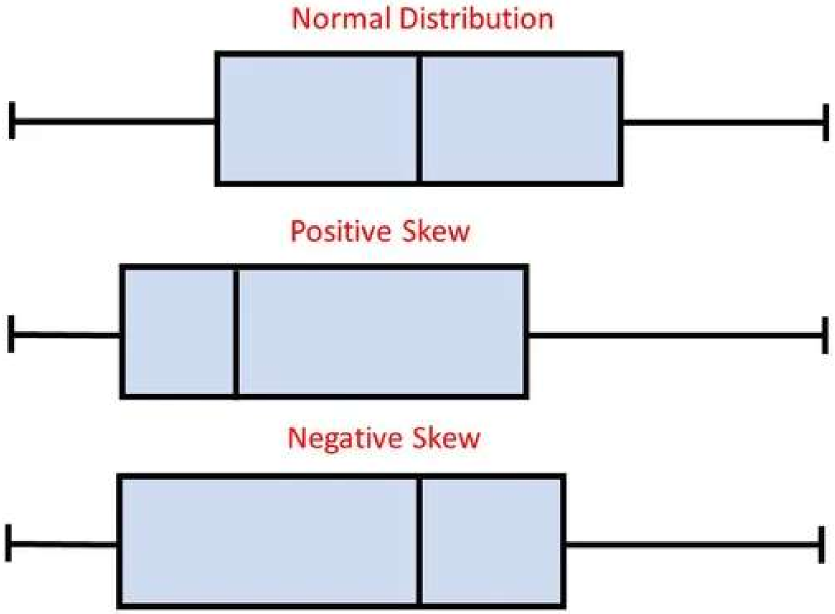
Box-and-Whisker Plots

- A **box-and-whisker plot is a graph that describes the shape of a distribution** in terms of the five-number summary:
  - minimum value
  - first quartile (25thpercentile)
  - median
  - third quartile (75th percentile)
  - maximum value
- The inner box shows the numbers that span the range from the first to the third quartile. A line is drawn through the box at the median.
- There are two “whiskers.”
  - One whisker is the line from the 25th percentile to the minimum value
  - the other whisker is the line from the 75th percentile to the maximum value

Box-and-Whisker Plots







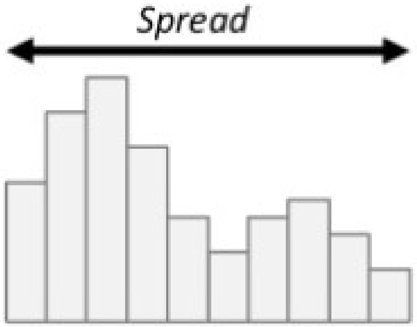
An outlier is an observation that is numerically distant from the rest of the data.

Extreme outliers are data points that are more extreme than Q1 - 3 * IQR or Q3 + 3 * IQR and marked with an asterisk (*) on the box plot.

Mild outliers are data points that are more extreme than Q1 - 1.5 * IQR or Q3 + 1.5 * IQR and are marked with a circle (O) on the box plot

Measures of Dispersion

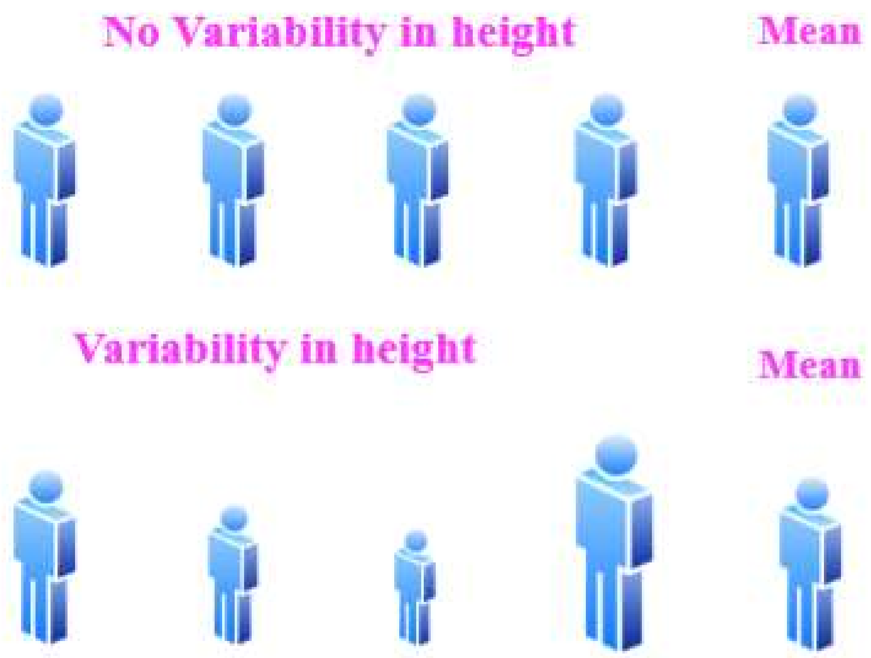
- The mean alone does not provide a complete or sufficient description of data
- The **Spread** of the data refers to how the data deviates from the position measure (whether it is the mean or the median)
- The degree to which the numerical data tend to spread about an average value is called the **variation or dispersion** of data
- It gives an indication of the amount of variation in the process
- It is an important indicator of quality
- It is important to have a measure of variability in order to control process variability and improve quality
- Remember that all manufacturing and transactional processes are variable to some degree



Measures of Dispersion

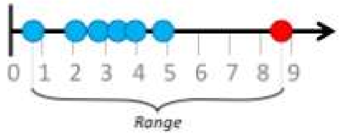
Common measures of Variability:

- Range
- Inter quartile range
- Mean absolute deviation
- Variance
- Standard deviation
- Z Scores
- Coefficient of variation



Range

- **Range is the difference between the largest and smallest observations.**
- It is the simplest measure of variability and usually denoted by ‘R’.
- It is good enough in many practical cases,
- however, it does not make full use of the available data.
- It can be misleading when the data is skewed or in the presence of outliers.
- Just one outlier will increase the range dramatically.



12 5 22 30 7 36 14 42 15 53 25 65

- Largest value=65
- Smallest Value=5
- Range= 65-5=60

Mean Absolute Deviation

- The mean absolute deviation of a dataset is the average distance between each data point and the mean
- It gives us an idea about the variability in a dataset

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

X_i	ith data point
$\bar{x}$	the mean of the data set
n	number of data points

Mean Absolute Deviation

Find the mean absolute deviation of the following data

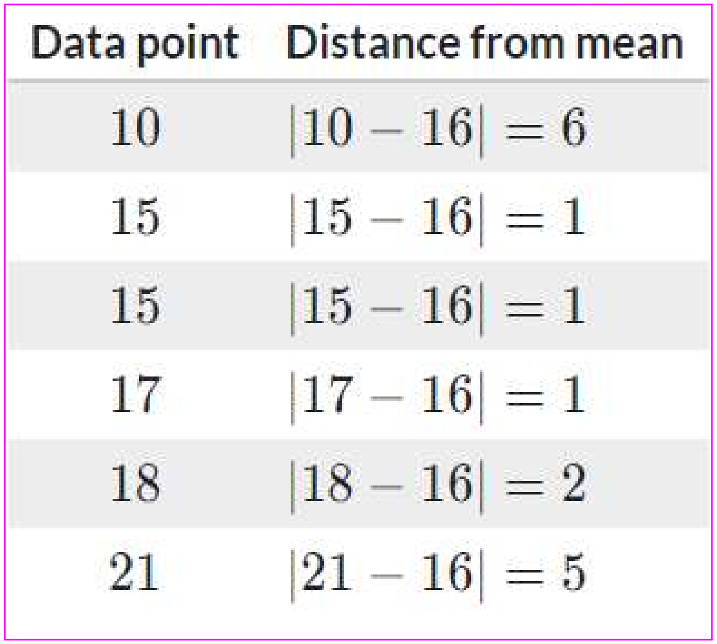
10 15, 15, 17, 18, 21

Mean is 96/6=16

Sum of absolute Deviation

$$6 + 1 + 1 + 1 + 2 + 5 = 16$$

$$\text{MAD} = \frac{16}{6} \approx 2.67$$



Data point	Distance from mean
10	$\lvert 10 - 16 \rvert = 6$
15	$\lvert 15 - 16 \rvert = 1$
15	$\lvert 15 - 16 \rvert = 1$
17	$\lvert 17 - 16 \rvert = 1$
18	$\lvert 18 - 16 \rvert = 2$
21	$\lvert 21 - 16 \rvert = 5$

Variance and Standard Deviation

- Range and inter quartile range measure the spread of data, both measures take into account only two of the data values.
- We need a measure that would *average the* total ($\sum$) distance between each of the data values and the mean.
- The variance is the average squared deviation from the population mean
- But for *all data sets, this* sum will *always equal zero because the mean is the center of the data.*
- *If the data value is* less than the mean, the difference between the data value and the mean would be negative (and distance is not negative).
- If each of these differences is squared, then each observation (both above and below the mean) contributes to the sum of the squared terms.
- The average of the sum of squared terms is called the **variance.**
- The **standard deviation,** which is the square root of variance, restores the data to their original measurement unit.
- The standard deviation measures the average spread around the mean.

Population variance and Population standard Deviation

- With respect to **variance, the *population variance σ^2, is the sum of the*** squared differences between each observation and the population mean divided by the population size, *N:*

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- the population *standard deviation, **σ**, is* the (positive) square root of the population variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Sample variance and Sample standard Deviation

The *sample variance, s*^{2}*, is the sum of the squared differences between each* observation and the sample mean divided by the sample size, *n, minus 1:*

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}$$

The sample standard deviation, s,

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}}$$

Variance and Standard Deviation of grouped data

Population	Sample
$\sigma^2 = \frac{\sum_{i=1}^{N} f(x_i - \mu)^2}{N}$ $\sigma = \sqrt[2]{\sigma}$	$s^2 = \frac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n - 1}$ $s = \sqrt[2]{s}$

Degrees of Freedom

- The degrees of freedom for a calculation is the number of values in the final calculation of a statistic/parameter that are free to vary.

Suppose if we draw 3 independent observations from a population where a population mean μ=8

i	xi	Xi-μ
1	9	9-8=1
2	4	4-8=-4
3	?	?

- Suppose if we have the same situation but μ is unknown and $\bar{x}$ is 5

i	xi	$\text{Xi-}\overline{\text{X}}$
1	9	9-5=4
2	4	4-5=-1
3	2	2-5=-3

One of the property of using sample mean is

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

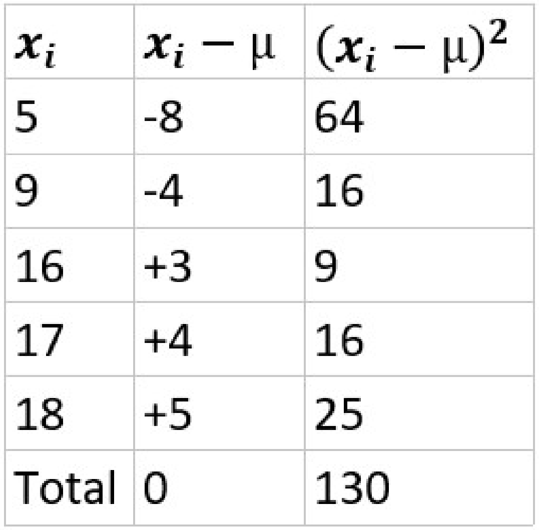
Degrees of freedom for sample variance is n-1 as it lost one observation

Degrees of Freedom

- When calculating the sample variance, one observation is already used. So We loss the freedom of choosing the last observation observation.
- Degrees of freedom for population variance is N, population parameter is fixed
- Sample mean may change depends upon the sample size and the population mean does not change
- We don’t loose any observation when we estimate the population variance
- the sample variance s^2 is an “unbiased estimator” of the population variance $\boldsymbol{\sigma^2}$ ***if the population variance is unknown***

Example: Population Variance and Standard Deviation

- Average of the squared deviations from the population mean



x_i	$x_i - \mu$	$(x_i - \mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
Total	0	130

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$= \frac{5+9+16+17+18}{5} = \frac{65}{5} = 13$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)}{N}$$

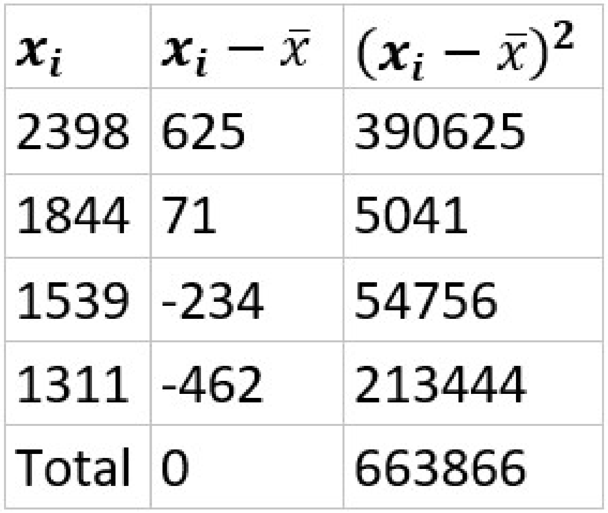
$$= \frac{130}{5} = 26$$

$$\sigma = \sqrt[2]{\sigma}$$

$$= \sqrt[2]{26} = 5.1$$

Example: Sample Variance Standard Deviation

Average of the squared deviations from the arithmetic mean



x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2398	625	390625
1844	71	5041
1539	-234	54756
1311	-462	213444
Total	0	663866

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

$$\bar{x} = \frac{2398+1844+1539+1311}{4} = \frac{7092}{4} = 1773$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$$

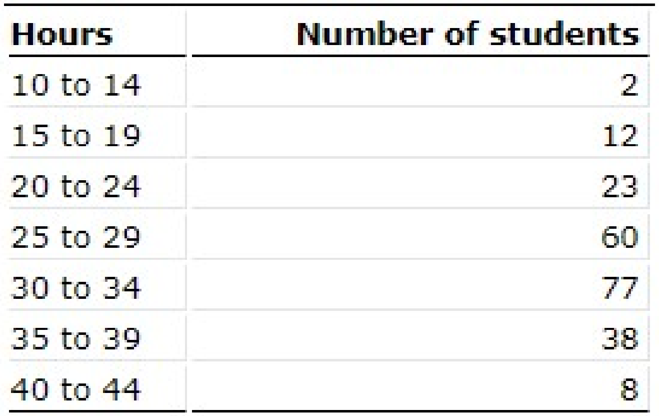
$$= \frac{663866}{3} = 221288.67$$

$$s = \sqrt[2]{s}$$

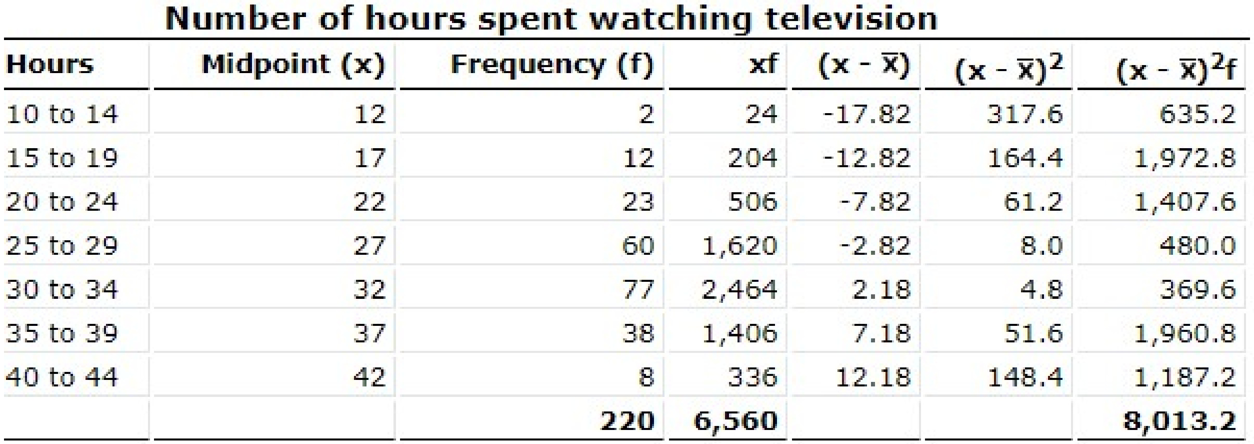
$$= \sqrt[2]{221288.67} = 470.41$$

Example: Sample Variance Standard Deviation grouped data

220 students were asked the number of hours per week they spent watching television. With this information, calculate the mean and standard deviation of hours spent watching television by the 220 students.



Hours	Number of students
10 to 14	2
15 to 19	12
20 to 24	23
25 to 29	60
30 to 34	77
35 to 39	38
40 to 44	8



Number of hours spent watching television

Hours	Midpoint (x)	Frequency (f)	xf	(x - $\bar{x}$)	(x - $\bar{x}$)^2	(x - $\bar{x}$)^{2}f
10 to 14	12	2	24	-17.82	317.6	635.2
15 to 19	17	12	204	-12.82	164.4	1,972.8
20 to 24	22	23	506	-7.82	61.2	1,407.6
25 to 29	27	60	1,620	-2.82	8.0	480.0
30 to 34	32	77	2,464	2.18	4.8	369.6
35 to 39	37	38	1,406	7.18	51.6	1,960.8
40 to 44	42	8	336	12.18	148.4	1,187.2
		220	**6,560**			**8,013.2**

$$\sigma^2 = \frac{\sum_{i=1}^{N} f(x_i - \mu)^2}{N}$$

$$\sigma = \sqrt[2]{\sigma}$$

$$= \sqrt{\frac{8{,}013.2}{220}}$$

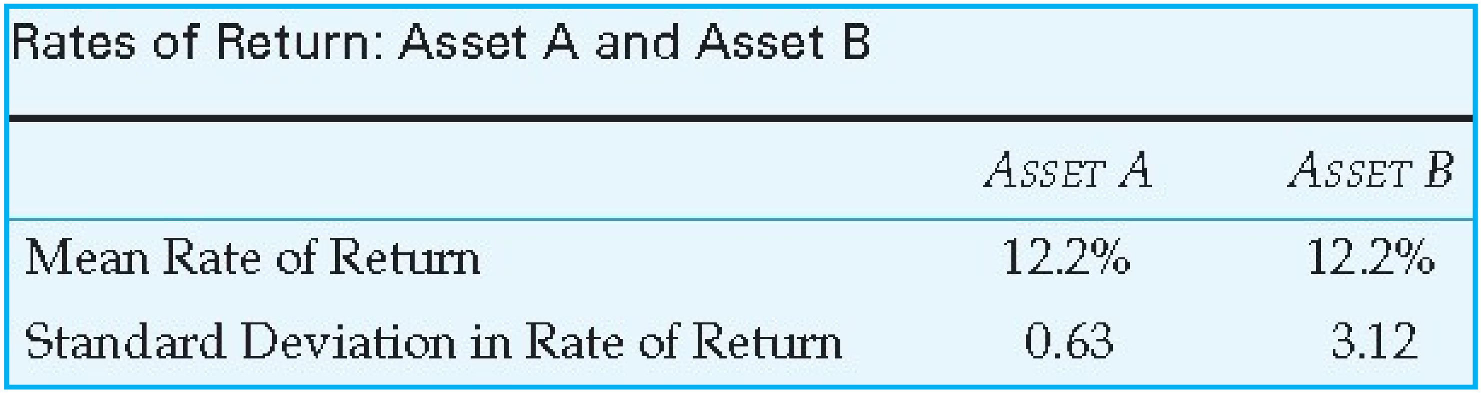
$$= \sqrt{36.42}$$

$$= 6.03$$

Applications of variance and standard deviation

- To Determine the spread of the data
- To determine the consistency of the variable. Eg. In the manufacture of fittings, such as nuts and bolts, the variation in the diameter must be small or the parts will not fit together
- To determine the number of data values that fall with in a specified interval in a distribution
- Indicator of financial risk
- Quality control
- Construction of quality control charts
- Process capability studies
- Comparing population
- House hold income in two cities
- Employee absenteeism in two cities

Comparing Risk of Two Assets with Equal Mean Rates of



Rates of Return: Asset A and Asset B

	Asset A	Asset B
Mean Rate of Return	12.2%	12.2%
Standard Deviation in Rate of Return	0.63	3.12

Since each asset has the same average rate of return of 12.2%, we can compare the standard deviations and determines that asset B is a more risky investment

Coefficient of Variation

- When the mean is not equal in the above example, We need to compare the coefficient of variation for both stocks rather than the standard deviations.
- The *coefficient of variation expresses* ***the standard deviation as a percentage of the mean*** (provided the mean is positive).
- The **coefficient of variation, CV, is a measure of relative dispersion**
- The *population coefficient of variation is*

$$\boxed{\text{CV} = \frac{\sigma}{\mu} \times 100\% \qquad \text{if } \mu > 0}$$

- The *sample coefficient of variation is*

$$\boxed{\text{CV} = \frac{s}{\bar{x}} \times 100\% \qquad \text{if } \bar{x} > 0}$$

Coefficient of Variation

- The table gives the prices of three securities at quarterly intervals. Which of the security is more consistent?

Security A	Security B	Security C
110	1463	10
115	1383	15
120	1194	25
185	1830	25
195	1934	22
120	1530	11
155	1464	24
230	1500	20
200	1500	12
190	1634	15
195	1440	18
190	1490	10

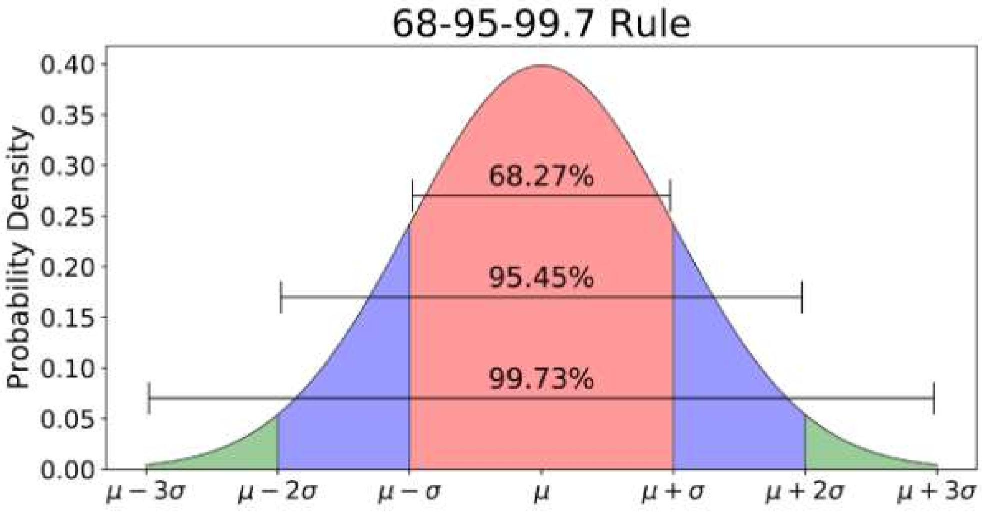
	Security A	Security B	Security C
Average	167.1	1530.2	17.3
S.D	41.0	195.1	5.9
C.V(%)	24.6	12.8	34

The security which has minimum CV is the one which is more consistent

Security B is more consistent ie CV(B)=12.8

Empirical Rule (68%, 95%, or Almost All)

- The empirical rule is an important rule of thumb that *is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed.*
- Approximately 68% of the observations are in the interval μ±1σ
- Approximately 95% of the observations are in the interval μ±2σ
- Almost all of the observations are in the interval μ±3σ



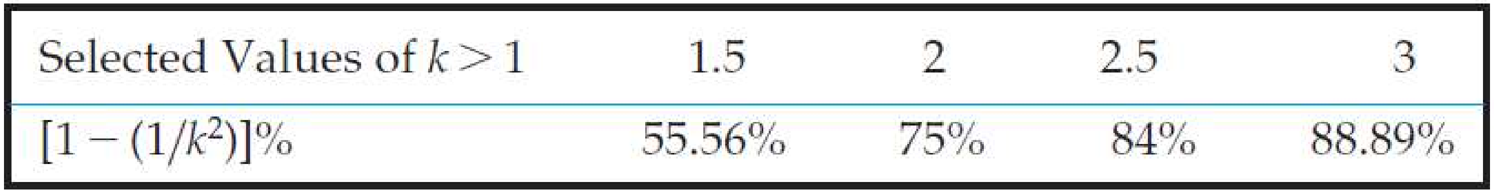
68% of the data is within 1 standard deviation, 95% is within 2 standard deviation, 99.7% is within 3 standard deviations

Chebyshev's Theorem

- A Russian mathematician, Pafnuty Lvovich Chebyshev (1821–1894), established data intervals for any data set, *regardless of the shape of the distribution.*
- For any population with mean μ, standard deviation σ, and *k >1, the percent* of observations that lie within the interval [μ±kσ]

$$\textit{at least } 100\left[1 - (1/k^2)\right]\%$$

- where k is the number of standard deviations.



Selected Values of $k > 1$	1.5	2	2.5	3
$[1-(1/k^2)]\%$	55.56%	75%	84%	88.89%

Lifetimes of Light bulbs

A company produces light bulbs with a mean lifetime of 1,200 hours and a standard deviation of 50 hours.

a. Describe the distribution of lifetimes if the shape of the population is unknown.

b. Describe the distribution of lifetimes if the shape of the distribution is known to be bell-shaped.

$$\begin{aligned}
\mu \pm 1\sigma &= 1{,}200 \pm 50 = (1{,}150, 1{,}250) \\
\mu \pm 2\sigma &= 1{,}200 \pm 2(50) = (1{,}100, 1{,}300) \\
\mu \pm 3\sigma &= 1{,}200 \pm 3(50) = (1{,}050, 1{,}350)
\end{aligned}$$

distribution is unknown	distribution is bell-shaped
we cannot make any conclusions about the percentage of bulbs that last between 1,150 hours and 1,250 hours. We can conclude that at least 75% of the lightbulbs will last between 1,100 hours and 1,300 hours and that at least 88.89% of the lightbulbs will last between 1,050 hours and 1,350 hours.	we can conclude that approximately 68% of the light bulbs will last between 1,150 hours and 1,250 hours; that approximately 95% of the light bulbs will last between 1,100 hours and 1,300 hours; and that almost all the bulbs will last between 1,050 hours and 1,350 hours.

z-Score

- That examines the location or position of a value relative to the *mean of the distribution*
- **z-score is a standardized value that indicates the number of standard deviations** a value is from the mean
- z-score greater than zero indicates that the value is greater than the mean
- z-score less than zero indicates that the value is less than the mean
- z-score of zero indicates that the value is equal to the mean
- If the data set is the entire population of data and the population mean μ, and the population standard deviation, σ, are known, then for each value, x_i, the corresponding z-score associated with x_i *is defined as follows:*

$$\boxed{z = \frac{x - \mu}{\sigma}}$$

- If the data set is the sample of data and the sample mean $\bar{x}$ and the sample standard deviation, s, are known, then for each value, x_i , the corresponding z-score associated with x_i *is defined as follows:*

$$\boxed{z = \frac{x - \bar{x}}{s}}$$

Example: Z score

- **Lifetimes of Light bulbs**

Consider the company discussed in previous example , which produces light bulbs with a mean lifetime of 1,200 hours and a standard deviation of 50 hours.

a. Find the z-score for a light bulb that lasts only 1,120 hours.

b. Find the z-score for a light bulb that lasts 1,300 hours.

Solution:

For life time 1120

$$\boxed{z = \frac{x_i - \mu}{\sigma} = \frac{1{,}120 - 1{,}200}{50} = -1.6}$$

For life time 1300

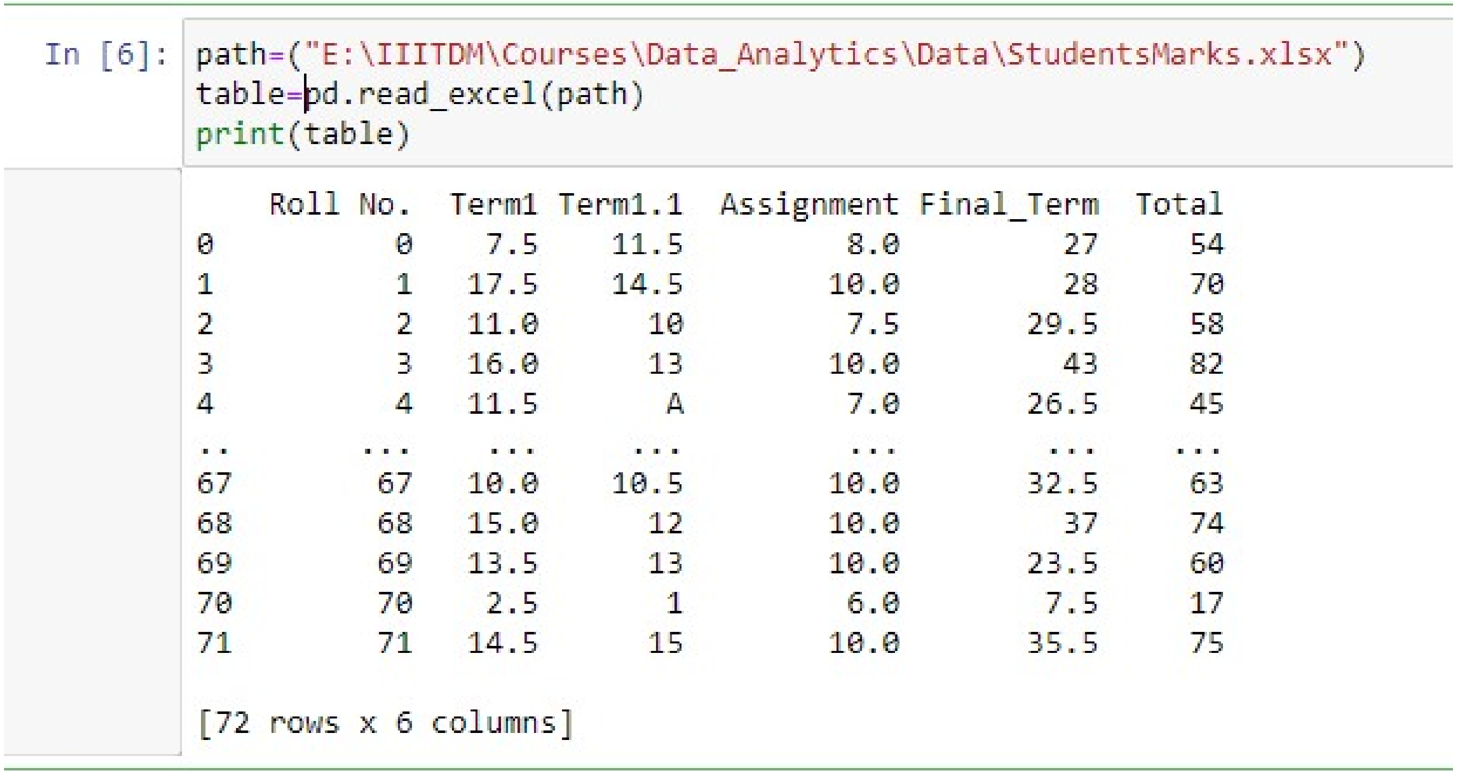
$$\boxed{z = \frac{x_i - \mu}{\sigma} = \frac{1{,}300 - 1{,}200}{50} = 2}$$

Describing a frequency distribution

- To describe the major characteristics of a frequency distribution , we need to calculate the following five quantities
- The total number of observations in the data set
- A measure of central tendency (mean, median and mode) that provides the information about the center or average value of the data set
- A measure of dispersion (variance, standard deviation) that indicates the spread of data
- A measure of skewness that shows the lack of symmetry in frequency distribution
- A measure of kurtosis that gives information about its peakedness
- It is interesting to note that all these quantities can be derived from the first four moments
  - First moment about the zero is the arithmetic mean
  - The second moment about mean is the variance
  - The third standardized moment is a measure of skewness
  - The fourth standardized moment is used to measure kurtosis

Measure of central tendency using python

Reading Table of student Marks



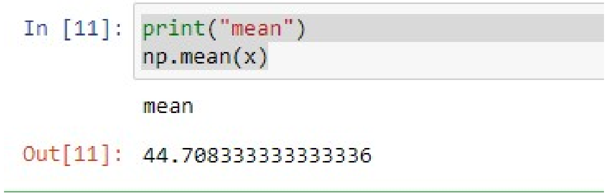
In [6]:

```
path=("E:\IIITDM\Courses\Data_Analytics\Data\StudentsMarks.xlsx")
table=pd.read_excel(path)
print(table)
```

```
    Roll No.  Term1 Term1.1  Assignment Final_Term  Total
0          0    7.5    11.5         8.0         27     54
1          1   17.5    14.5        10.0         28     70
2          2   11.0      10         7.5       29.5     58
3          3   16.0      13        10.0         43     82
4          4   11.5       A         7.0       26.5     45
..       ...    ...     ...         ...        ...    ...
67        67   10.0    10.5        10.0       32.5     63
68        68   15.0      12        10.0         37     74
69        69   13.5      13        10.0       23.5     60
70        70    2.5       1         6.0        7.5     17
71        71   14.5      15        10.0       35.5     75

[72 rows x 6 columns]
```

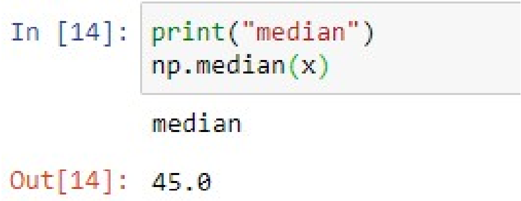
Measure of central tendency using python

- **Mean**



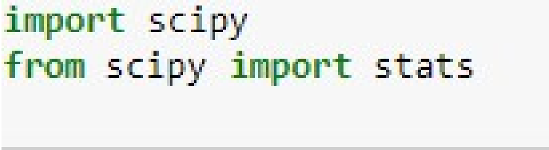
```
In [11]: print("mean")
         np.mean(x)

         mean

Out[11]: 44.708333333333336
```

- **Median**

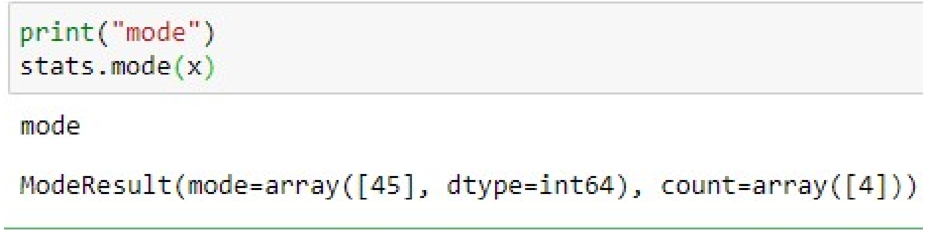


```
In [14]: print("median")
         np.median(x)

         median

Out[14]: 45.0
```

- **Mode**



```
import scipy
from scipy import stats
```

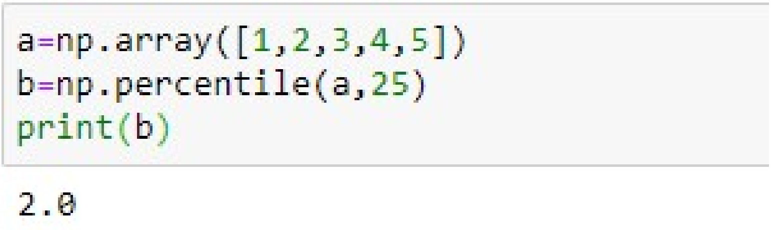


```
print("mode")
stats.mode(x)

mode

ModeResult(mode=array([45], dtype=int64), count=array([4]))
```

Number 45 occurs 4 time

Measure of central tendency using python

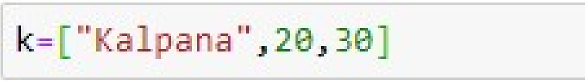
- Percentile



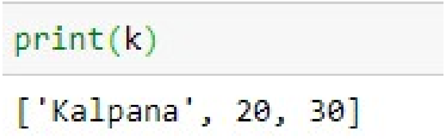
```
a=np.array([1,2,3,4,5])
b=np.percentile(a,25)
print(b)
```

```
2.0
```

- For Loop

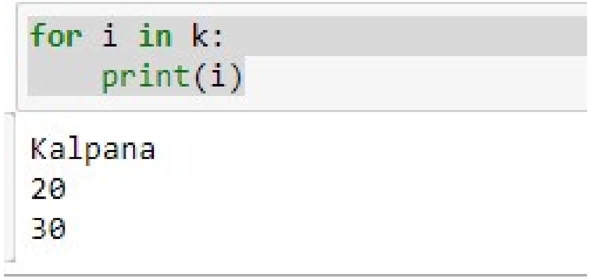


```
k=["Kalpana",20,30]
```



```
print(k)
```

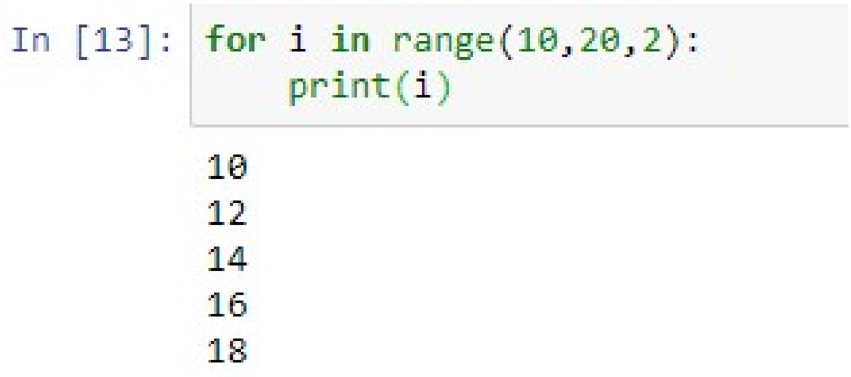
```
['Kalpana', 20, 30]
```



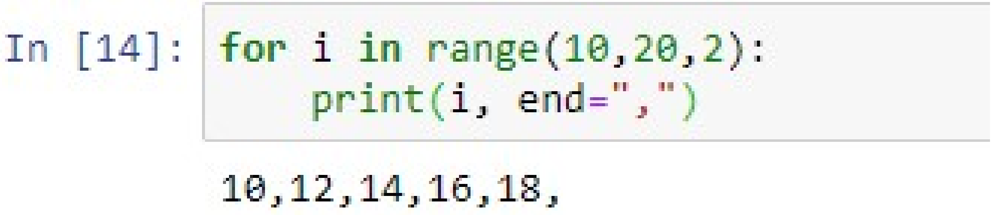
```
for i in k:
    print(i)
```

```
Kalpana
20
30
```

For loop in python

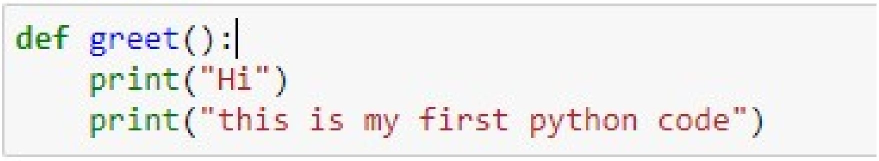


```
In [13]: for i in range(10,20,2):
             print(i)

10
12
14
16
18
```

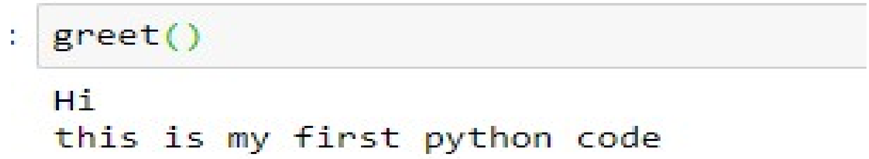


```
In [14]: for i in range(10,20,2):
             print(i, end=",")

10,12,14,16,18,
```

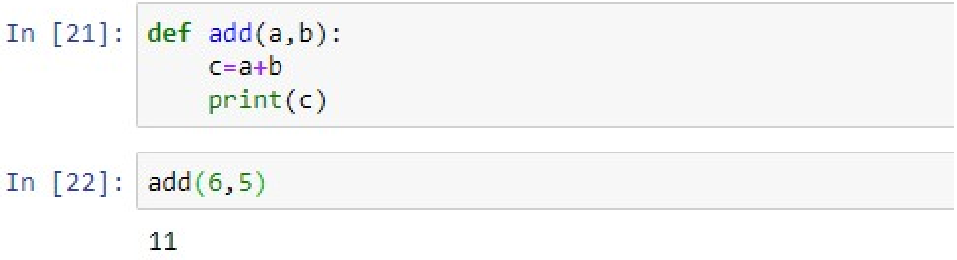
Functions in python



```
def greet():
    print("Hi")
    print("this is my first python code")
```

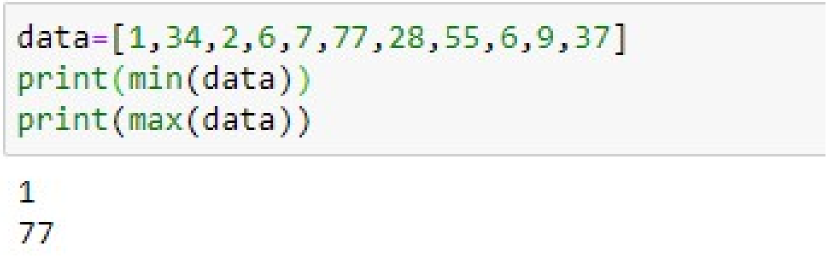


```
: greet()

Hi
this is my first python code
```



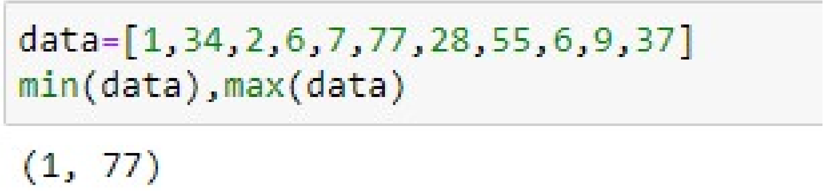
```
In [21]: def add(a,b):
             c=a+b
             print(c)

In [22]: add(6,5)

         11
```

Finding Minimum and maximum value



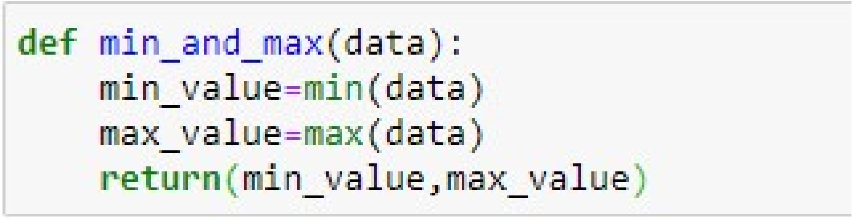
```
data=[1,34,2,6,7,77,28,55,6,9,37]
print(min(data))
print(max(data))
```

```
1
77
```

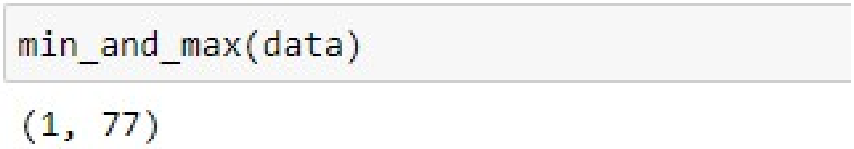


```
data=[1,34,2,6,7,77,28,55,6,9,37]
min(data),max(data)
```

```
(1, 77)
```



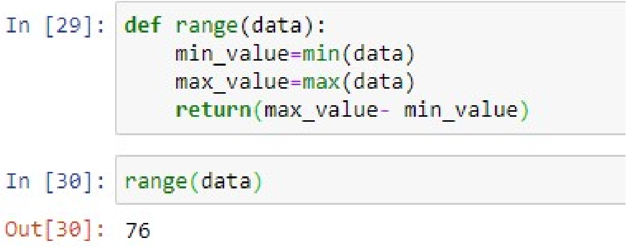
```
def min_and_max(data):
    min_value=min(data)
    max_value=max(data)
    return(min_value,max_value)
```



```
min_and_max(data)
```

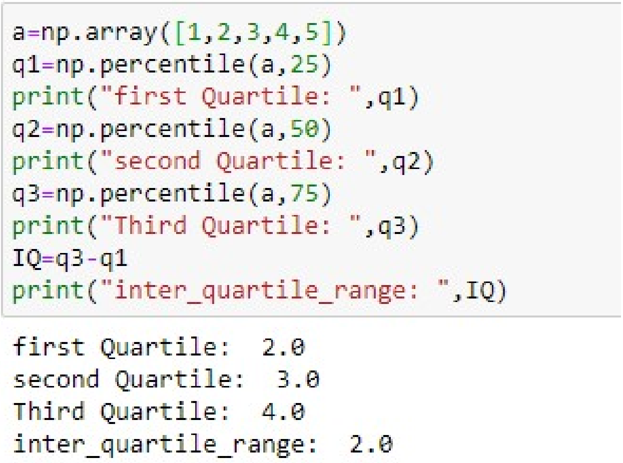
```
(1, 77)
```

To find Range



```
In [29]: def range(data):
             min_value=min(data)
             max_value=max(data)
             return(max_value- min_value)

In [30]: range(data)

Out[30]: 76
```

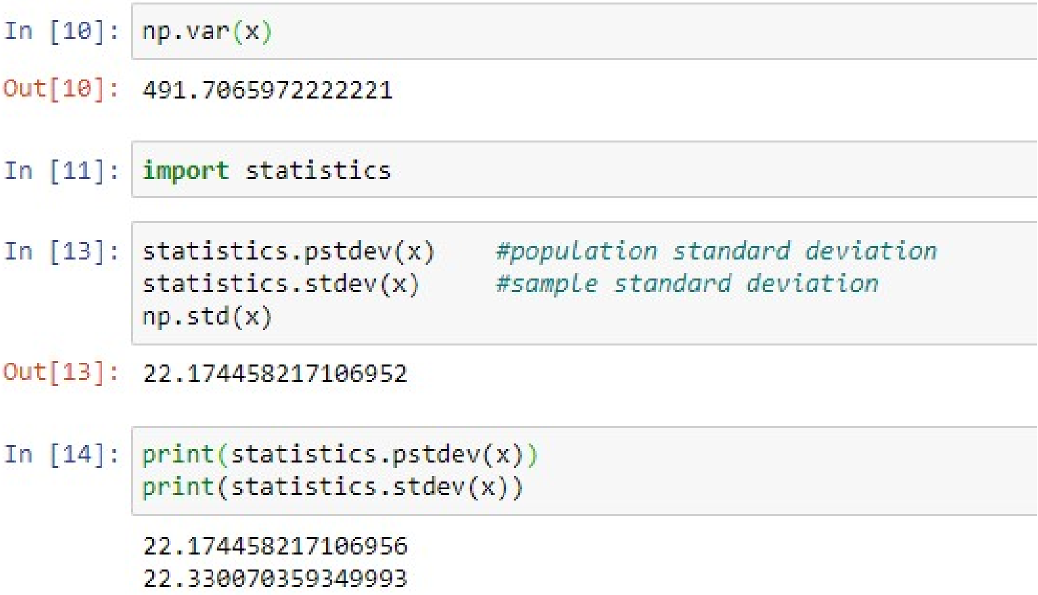
To find quartile



```
a=np.array([1,2,3,4,5])
q1=np.percentile(a,25)
print("first Quartile: ",q1)
q2=np.percentile(a,50)
print("second Quartile: ",q2)
q3=np.percentile(a,75)
print("Third Quartile: ",q3)
IQ=q3-q1
print("inter_quartile_range: ",IQ)
```

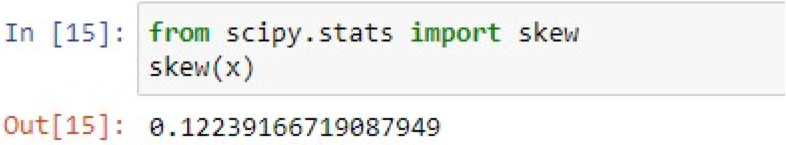
```
first Quartile:  2.0
second Quartile:  3.0
Third Quartile:  4.0
inter_quartile_range:  2.0
```

Variance and standard deviation



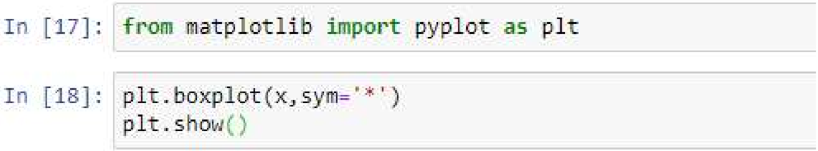
```
In [10]: np.var(x)

Out[10]: 491.7065972222221

In [11]: import statistics

In [13]: statistics.pstdev(x)    #population standard deviation
         statistics.stdev(x)     #sample standard deviation
         np.std(x)

Out[13]: 22.174458217106952

In [14]: print(statistics.pstdev(x))
         print(statistics.stdev(x))

         22.174458217106956
         22.330070359349993
```

To find Skew

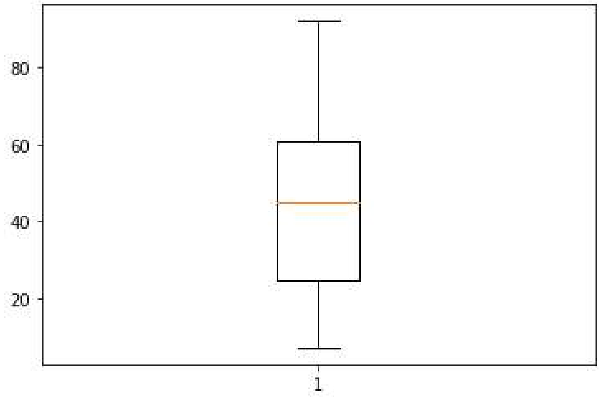


```
In [15]: from scipy.stats import skew
         skew(x)

Out[15]: 0.12239166719087949
```

Box Plot



```
In [17]: from matplotlib import pyplot as plt

In [18]: plt.boxplot(x,sym='*')
         plt.show()
```



Measures of Relationships between Variables

- Numerical ways to describe a linear relationship
  - *Covariance*
  - *Correlation*
- **Covariance**
  - A measure of the linear relationship between two variables
  - A positive value indicates a direct or increasing linear relationship
  - A negative value indicates a decreasing linear relationship

✓This value varies if the two variables have different units.

✓It does not provide a measure of the strength of the relationship between two variables

Population covariance	Sample covariance
$Cov(x,y) = \sigma_{xy} = \dfrac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$	$Cov(x,y) = s_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

where x_i *and* y_i *are the observed values,* μ_x *and* μ_y *are the population means, N is the population size, n sample size, and* $\bar{x}$ and $\bar{y}$ *are the sample means*

Measures of Relationships between Variables

- **Correlation Coefficient**
  - A standardized measure of the linear relationship between two variables
  - It is generally a more useful measure because it provides both the *direction and the strength of a relationship*
  - The covariance and corresponding correlation coefficient have the same sign (both are positive or both are negative).
  - Computed by dividing the covariance by the product of the standard deviations of the two variables
- A population correlation coefficient, ρ, is

$$\boxed{\rho = \frac{Cov(x, y)}{\sigma_x \sigma_y}}$$

- A sample correlation coefficient, r, is

$$\boxed{r = \frac{Cov(x, y)}{s_x s_y}}$$

- A useful rule to remember is that a relationship exists if

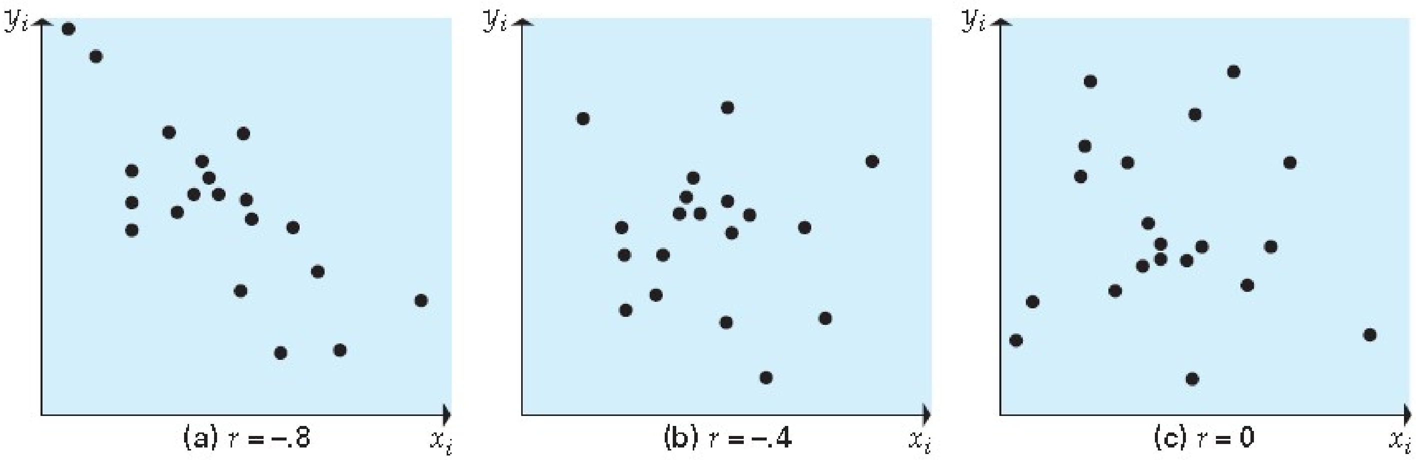
$$\boxed{|r| \geq \frac{2}{\sqrt{n}}}$$

Correlation Coefficient

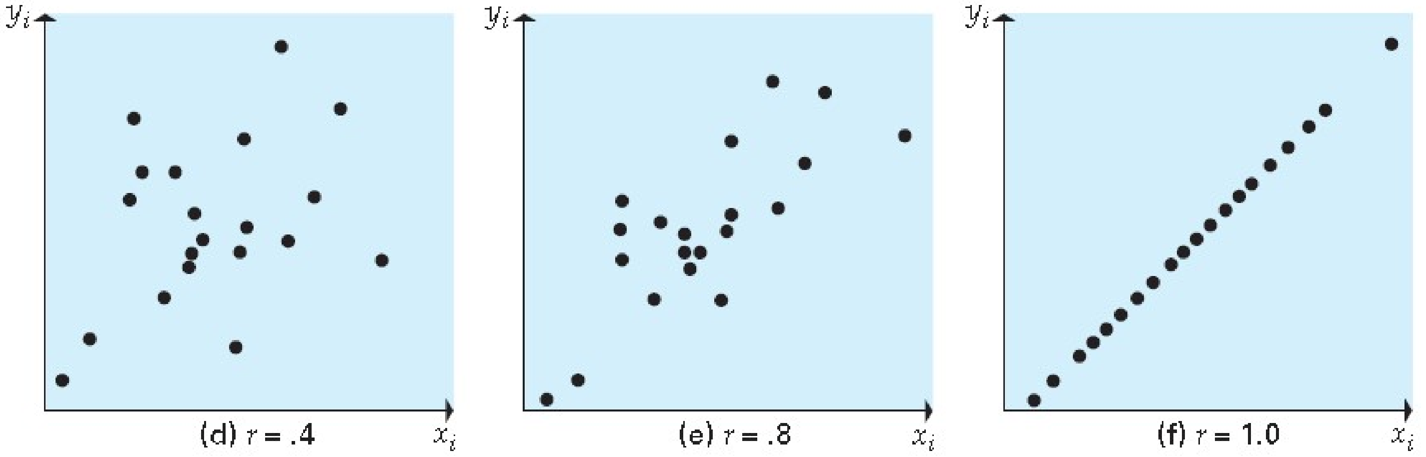
➢The correlation coefficient ranges from -1 to +1.

- The closer r is to +1, the closer the data points are to an increasing straight line, indicating a positive linear relationship.
- The closer r is to -1, the closer the data points are to a decreasing straight line, indicating a negative linear relationship.
- When r = 0, there is no linear relationship between x and y—but not necessarily a lack of relationship.

Sales Vs Time



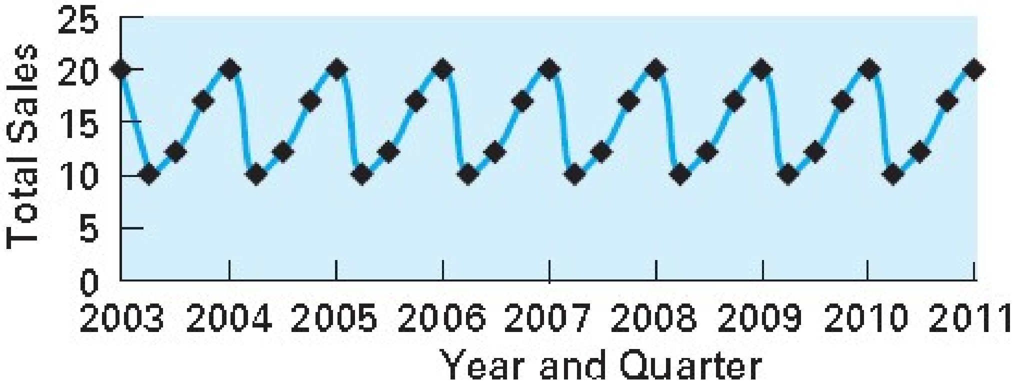
Correlation Coefficient



(d) $r = .4$

(e) $r = .8$

(f) $r = 1.0$



Example

Facebook Posts (site updates) and Fan Interactions

Facebook posts (updates), x	16	31	27	23	15	17	17	18	14
Fan interactions, y	165	314	280	195	137	286	199	128	462

The mean and the variance in the number of Facebook posts are found to be approximately

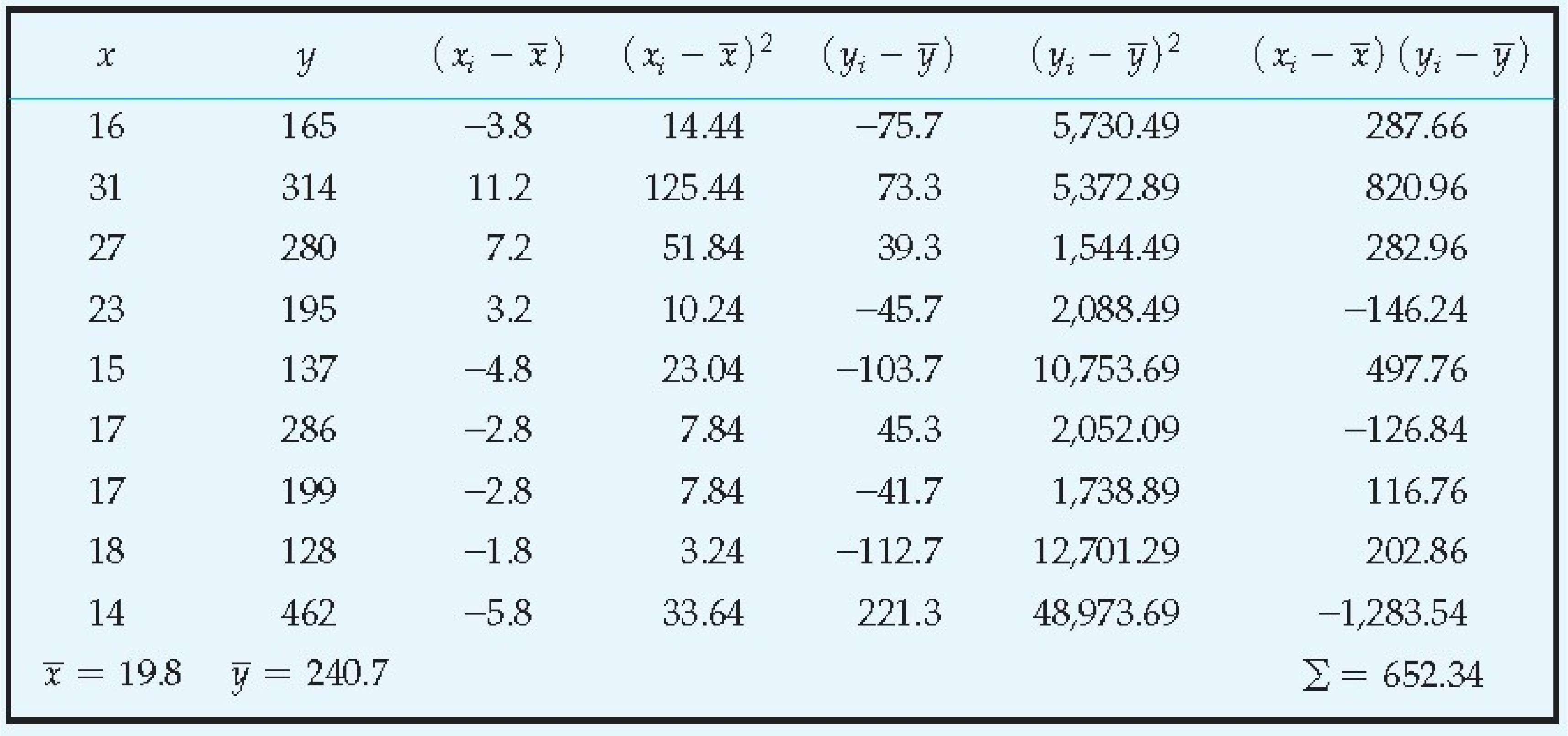
$$\bar{x} = 19.8 \quad \text{and} \quad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} = 34.694$$

The mean and the variance in the number of fan interactions are found to be approximately

$$\bar{y} = 240.7 \quad \text{and} \quad s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1} = 11{,}369.5$$

Example

Facebook Posts and Fan Interactions (Covariance and Correlation)



x	y	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
16	165	–3.8	14.44	–75.7	5,730.49	287.66
31	314	11.2	125.44	73.3	5,372.89	820.96
27	280	7.2	51.84	39.3	1,544.49	282.96
23	195	3.2	10.24	–45.7	2,088.49	–146.24
15	137	–4.8	23.04	–103.7	10,753.69	497.76
17	286	–2.8	7.84	45.3	2,052.09	–126.84
17	199	–2.8	7.84	–41.7	1,738.89	116.76
18	128	–1.8	3.24	–112.7	12,701.29	202.86
14	462	–5.8	33.64	221.3	48,973.69	–1,283.54
$\bar{x} = 19.8$	$\bar{y} = 240.7$					$\sum = 652.34$

Example

$$Cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n - 1} = \frac{652.34}{8} = 81.542$$

$$r = \frac{Cov(x,y)}{s_x s_y} = \frac{81.542}{\sqrt{34.694}\sqrt{11{,}369.5}} = 0.1298$$

$$|0.1298| < \frac{2}{\sqrt{9}} = 0.67$$

We conclude that there is not sufficient data to think that there is a strong linear relationship between Facebook posts and fan interaction.

Reference

- statistics-for-business-and-economics-8th-edition, Paul newbold, William L.Carlson,Betty M.Thome
- Business statistics for contemporary Decision making , 6th edition by Ken Black
- python/https://medium.com/budding-data-scientist
- statistics-for-business-and-economics-11th edition, David R. Anderson,Dennis J. Sweeney,Thomas A. Williams

thank you
danke
謝謝
gracias
obrigado
merci
ngiyabonga
teşekkür ederim
рахмат
Баярлалаа
спасибо
dank je
bedankt
dziękuję
hvala
mauruuru
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
tapadh leat
arigatò
takk
dakujem
grazie
terima kasih
sagolun
ありがとう
감사합니다
мерси
ευχαριστώ
enkosi
kiitos
dankie
dhanyavad
faafetai lava
mersi
vinaka
blagodaram
welalin
tack
misaotra
paldies
grazzi
хвала
asante
manana
obrigada
mamnun
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
trugarez
chnorakaloutioun
gracies
sulpay
taiku
sobod
dekuji
mesi
najis tuke
kam-sah hamnida
xiexie
nanni
bayarlalaa
grazie
kiitos
dankon
aciu
நன்றி