Data representation

Concept of Statistic

- ***Statistics*** **is defined as a collection of methods for collecting, organizing and representing data**

This mean the following:

- 1- Collect Data (Raw Data)
- 2- Organized Data (Table- Graph)
- 3- Analysed Data
- 4- Interpret the Data (take Decision)

There two types of statistics:

وصفيDescriptive statistics :

which describe data or represent it, data can be described using frequency count, average, modes, median and standard deviation

إستنتاجيInferential statistics:

trying to guess the population parameters using the given sample . it performs hypothesis testing and mares prediction .

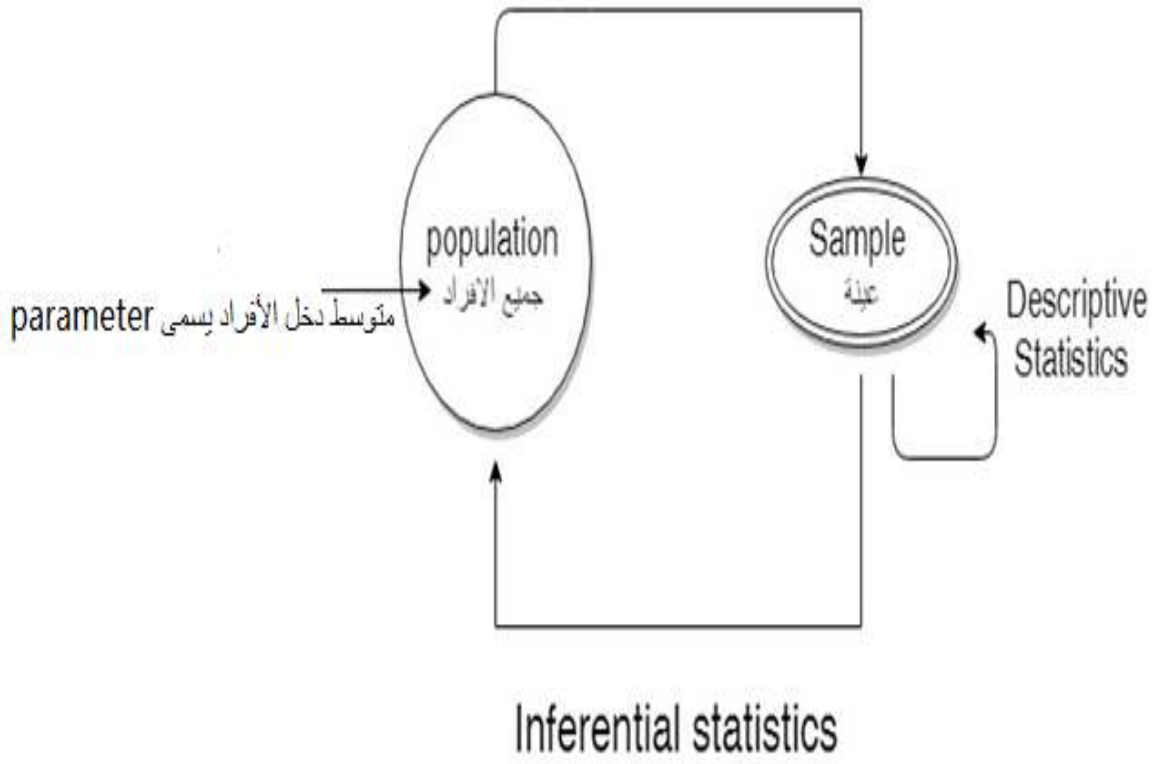


Figure 1.1 shows the scope of statistics, which includes population, sample, parameters and statistic.

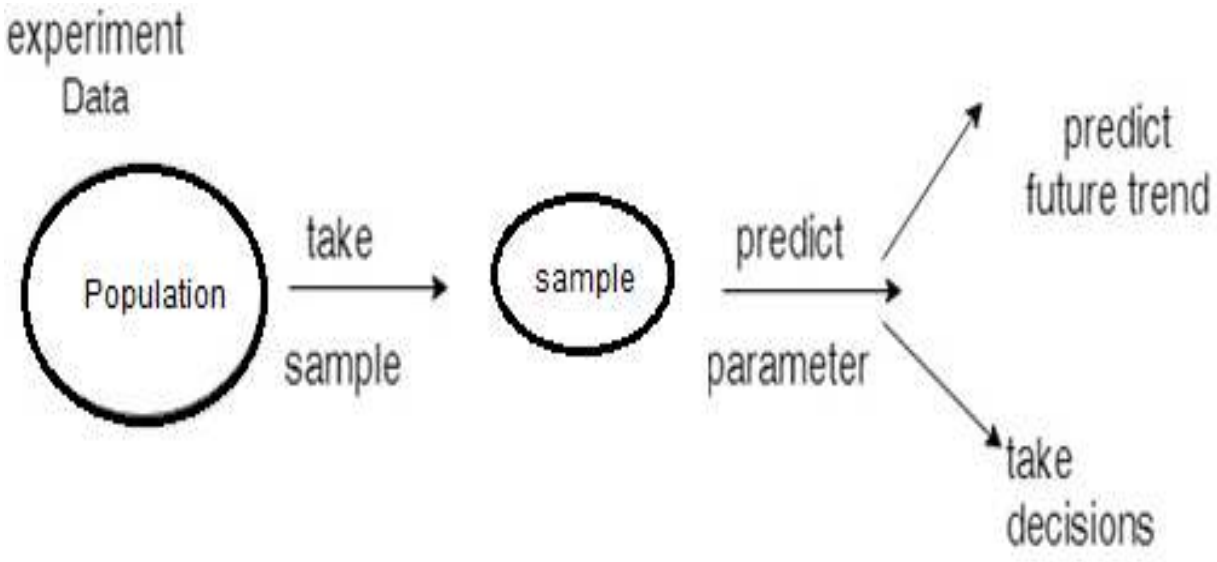


Figure 1.2: stages of making statistical experiments

Some Important Definitions

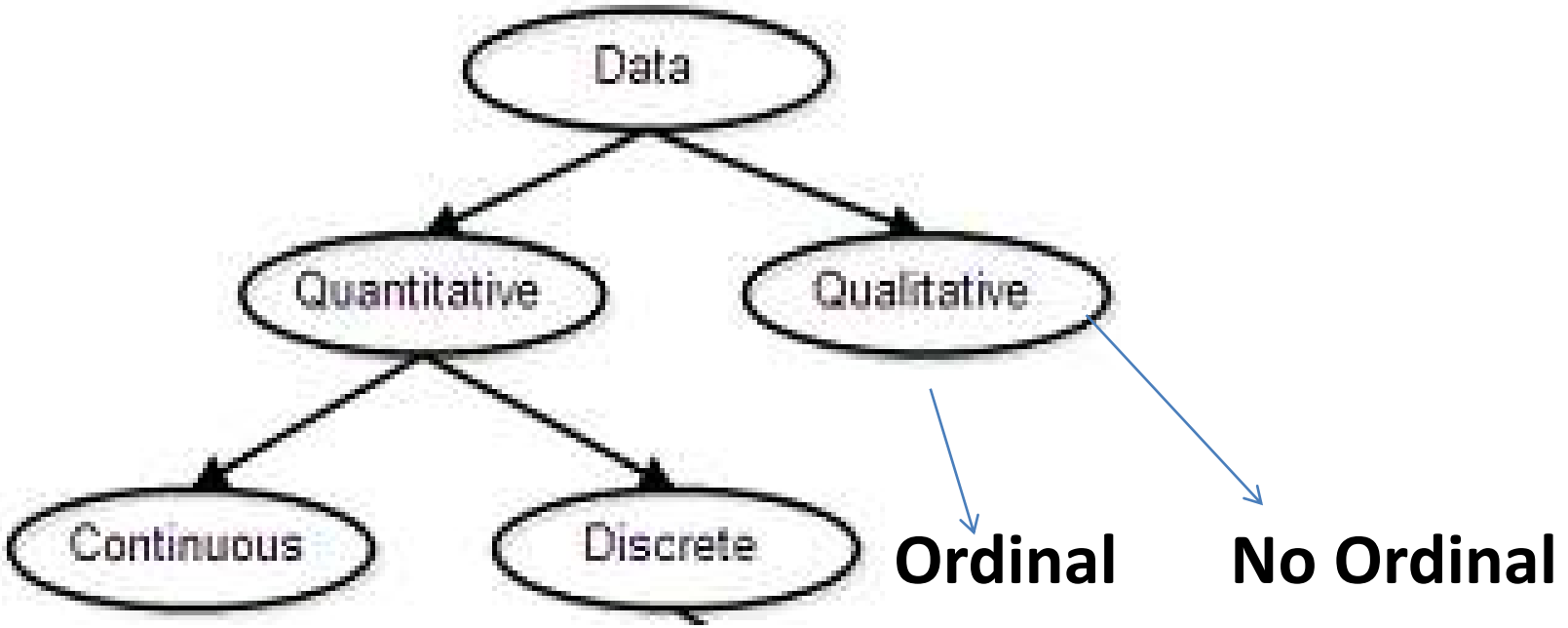
- The population : the collection all elements (scores, persons, measurements, …. Etc.) to be studied
- Sample : subset of the population
- Parameter : characteristic or measure obtained from the population .
- Statistic : is a characteristic or measure obtained from the sample .

Types of Data

The data can be:

Qualitative use properties or characteristic of Data such as “female”, “male”, or “Egyptian”, “Saudi” etc. generally qualitative data can’t be represented with numbers can be **Ordinal** or non-**Ordinal**.

Quantitative data is the data that can be represented numerically such as 0, 5, 21, 300, 0.12, etc.
Quantitative data can be **discrete** or **continuous**


Data
Quantitative
Qualitative
Continuous
Discrete
Ordinal
No Ordinal

Discrete Data : The data that its value is undividable such as
Number of children in a family
Number of students
Number of cars sold in a day
Generally , we wouldn't expect to find 2.5 children or $\frac{1}{2}$ of a car sold in a day.

Binary date is the data that can have one of two values, such as of 1 or 0, Binary data is discrete data type.

Continuous Data : the observation that can take any value within finite or infinite interval real numbers, in other words it contains fractions.
Examples are :
Weight Height of student Time to run 500 meters Age
The salary of employees

Type of Variables

- 1- **Dependant Variable**
- Variable that are dependent in each others
- Example: the pressure of blood for ill people before and after take the doss

- 2- **Independent Variable**
- Not dependent to each others

Example: **Degree** of student in deferent university in **statistic courses**

Organization of Data

The statistician must organize data into meaningful way and present it so readers Can understand the distribution of data .

Frequency table

Counting the number of repetition of specific value in the data.

Example : draw frequency table for the.

1 2 6 7 2 6 5 7 1 5 5 6 5 5 2

Solution :

Date	1	2	5	6	7
frequency	2	3	5	3	2

Grouped Frequency table

Counting the frequency of repeating a value in specific range

Example : Draw the grouped frequency table of the following data with 5 classes .

12,22,18,9,25,31,28,19,22,27,32,14

Solution :

Find the min and max values min 9, max 32 .

wide the range over the class number

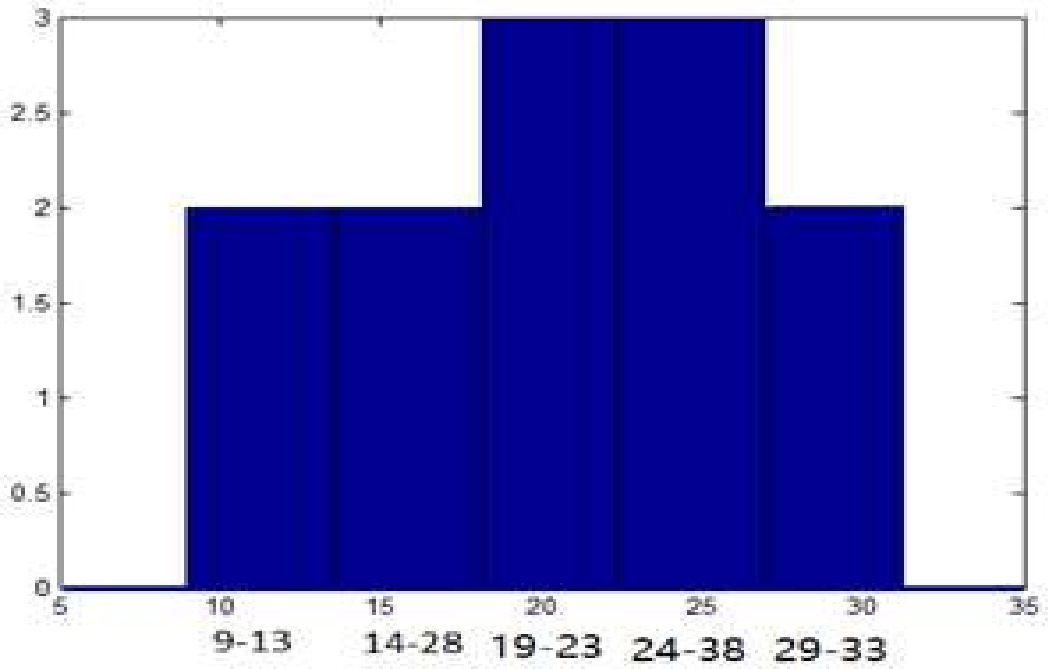
$$\frac{32-9}{5} = \frac{range}{\#\ classes} = \frac{23}{5} = 4.6$$

Class	Frequency
9 – 13	2
14 – 18	2
19 – 23	3
24 – 28	3
29 – 33	2

Histogram المدرج التكراري

Displaying the frequency table with touching and non-overlapping bars .

Example : Use histogram to represent the above grouped frequency table



Frequency histogram with MATLAB use the following commands

```
>> a=[12,22,18,9,25,31,28,19,22,27,32,14]
>> hist ( a,5 )
```

Measures of Location (Central Tendency)

مقاييس_النزعة_المركزية

Measures of Central Tendency

- The data (observations) often tend to be concentrated around the center of the data.
- Some measures of location are: the mean, median and mode.
- These measures are considered as representatives (or typical values) of the data.

Measures of Central Tendency

- They are designed to give some quantitative measures of where the center of the data is in the sample.

The Sample Mean:

- Is the most common measure of central tendency
- The sum of the values (positive , negative or zero) divided by the number of values
- Is called the Mean , Sample Mean , Arithmetic Mean and average.

The Sample Mean:

If $X_1, X_2, \ldots\ldots\ldots, X_n$ are the sample values, then the sample mean is:

$$\overline{X} = \frac{X_1 + X_2 + \ldots\ldots\ldots + X_n}{n}$$

Using summation:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The Sample Mean (Example 1):

Suppose that the following sample represents the ages (in year) of a sample of 3 men: $X_1 = 30,\ X_2 = 35,\ X_3 = 27.$

Then, the sample mean is:

$$\overline{X} = \frac{30+35+27}{3} = \frac{92}{3} = 30.67$$

The Sample Mean (Example 2):

- **For what value of X will 8 and X have the same sample mean as 27 and 5?**

Solution:

First, find the mean of 27 and 5:

$$\frac{27+5}{2}=16$$

Now, find the X value, knowing that the sample mean of X and 8 must be 16 :

$$\frac{X+8}{2}=16$$

cross multiply and solve: $32 = X + 8$ → $X = 24$

The Sample Mean (Example 3):

- **On his first 5 Stat. tests, Omar received the following marks : 72, 86, 92, 63, and 77. What test mark must Omar earn on his sixth test so that his average for all six tests will be 80? .**

- **Solution**

Set up an equation to represent the situation.

$$\frac{72+86+92+63+77+X}{6}=80 \longrightarrow \quad X=90$$

Omar must get a 90 on the sixth test.

The Sample Mean (Advantages):

- **Most popular measure** in fields such as business, engineering and computer science.
- **It is unique** - there is only one answer.
- **Useful when comparing** sets of data.

The Sample Mean (Disadvantages):

- **Affected by extreme values (outliers)**

Example.

- **The sample mean of 2,3,4 is 3.**
- **The sample mean of 2,3,40 is 15.**
- **The mean increased from 3 to 15 because 40 is an extreme value.**

The following MATLAB command can be used to get the arithmetic mean

```
>> mean(a)
```

The Sample Median

- The purpose of the sample median is to reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliers

- The value which divides the data into two equal halves, with half of the data being lower than the median and half higher than the median.

The Sample Median (Steps):

If $X_1, X_2, \ldots\ldots\ldots, X_n$ are the sample values, then the sample median computed as follows:

- Sort the values into ascending order.
- If we have an odd number (*n*) of values, the median is the middle value.

$$M_e = X_{\frac{1}{2}(n+1)}$$

The Sample Median (Steps):

- If we have an **even number** of values, the median is the sample **mean** of the two middle values.

$$M_e = \frac{1}{2}\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1} \right)$$

The Sample Median (Example 1):

Compute The sample median of

(12, 24, 19, 20, 7) .

Solution

- **Sort the values into ascending order.**

7,12, 19, 20, 24

- **Number of values is 5 is an odd the sample median is :**

$$M_e = X_{\frac{1}{2}(n+1)} = X_{\frac{1}{2}(5+1)} = X_3 = 19$$

The Sample Median (Example 2):

Compute The sample median of

(12, 24, 19, 20, 7 , 5) .

Solution

- **Sort the values into ascending order.**

5,7,12, 19, 20, 24

- **Number of values is 6 is an even the sample median is :**

$$M_e = \frac{1}{2}\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1} \right) = \frac{1}{2}\left(X_{\frac{6}{2}} + X_{\frac{6}{2}+1} \right)$$

$$= \frac{(X_3 + X_4)}{2} \qquad = \frac{(12+19)}{2} \qquad = 15.5$$

The Sample Median (Advantages):

- **Extreme values do not affect the median as strongly as they do to the mean.**
- **It is unique - there is only one answer.**
- **Useful when comparing sets of data.**

The Sample Median (Disadvantages):

Not as popular as sample mean

The following MATLAB command can be used to get the median

≫ a=[12,22,18,9,25,31,28,19,22,27,32,14]

≫ **med = median (a)**

The Sample Mode :

Mode measures the data which occurs most frequently

Example : the set $2, 4, 9, 8, 8, 5, 3$ has mode of 8

Example : the set $2, 2, 9, 8, 8, 5, 3$ has mode of 2, 8

No sample mode of the list (1, 6, 2, 7, 3, 5).

The Sample Mode (Advantages):

- Extreme values do not affect the mode.

Example:

- The sample mode of the list (1, 2, 2, 3, 3, 3, 4) is 3.
- The sample mode of the list (1, 2, 2, 3, 3, 3, 4000) is 3.

The Sample Mode (Disadvantages):

- **Not as popular as mean and median.**
- **Not necessarily unique - may be more than one answer**
- **When no values repeat in the data set, the mode is every value and is useless.**

The following MATLAB command can be used to get the mode

≫ a=[12,22,18,9,25,31,28,19,22,27,32,14]

≫ $\mathbf{m} = \mathbf{m}\boldsymbol{ode}\ (\mathbf{a})$

Quartiles

The three values that split a set of ranked data values for a variable into four equal parts — quarters, or quartiles.

- The **First quartile (Q_1)** is the value such that 25% of the ranked data are smaller and 75% are larger.

$$Q_1 = \begin{cases} x_{\frac{(n+1)}{4}} & \text{if } n \text{ is Odd} \\ \frac{1}{2}\left(x_{\frac{n}{4}} + x_{\frac{n}{4}+1} \right) & \text{if } n \text{ is Even} \end{cases}$$

- The **Second quartile (Q_2)** is another name for the median.
- The **Third quartile (Q_3)** is the value such that 75% of the ranked data are smaller and 25% are larger.

$$Q_3 = \begin{cases} x_{\frac{3(n+1)}{4}} & \text{if } n \text{ is Odd} \\ \frac{1}{2}\left(x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1} \right) & \text{if } n \text{ is Even} \end{cases}$$

Example 4: Find the median, first quartile, third quartile of the following data of scores:
12 25 15 5 22 7 14 36 53 30 42

Solution: First, arrange the data in ascending order:

5 7 12 14 15 22 25 30 36 42 53

Q_1 *Median* Q_3

$$Q_1 = x_{\frac{(n+1)}{4}} = x_{\frac{(11+1)}{4}} = x_3 = 12 \qquad Q_3 = x_{\frac{3(n+1)}{4}} = x_{\frac{3(11+1)}{4}} = x_{\frac{36}{4}} = x_9 = 36$$

Example 5: Find the median, first quartile, third quartile, if we add 65 to Example4.
Solution: First, arrange the data in ascending order:

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

Q_1 *Median* Q_3

$$Median(Q_2)=\frac{1}{2}\left(x_{\frac{2n}{4}}+x_{\frac{2n}{4}+1}\right)=\frac{1}{2}\left(x_{\frac{2*12}{4}}+x_{\frac{2*12}{4}+1}\right)=\frac{1}{2}(x_6+x_7)=\frac{1}{2}(22+25)=23.5$$

$$Q_1=\frac{1}{2}\left(x_{\frac{n}{4}}+x_{\frac{n}{4}+1}\right)=\frac{1}{2}\left(x_{\frac{12}{4}}+x_{\frac{12}{4}+1}\right)=\frac{1}{2}(x_3+x_4)=\frac{1}{2}(12+14)=13$$

$$Q_3=\frac{1}{2}\left(x_{\frac{3n}{4}}+x_{\frac{3n}{4}+1}\right)=\frac{1}{2}\left(x_{\frac{3*12}{4}}+x_{\frac{3*12}{4}+1}\right)=\frac{1}{2}(x_9+x_{10})=\frac{1}{2}(36+42)=39$$

When the result of this arithmetic is not a whole number (whether *n* is odd or even) use the following formulas for Q_1 and Q_3 :

$$Q_1=x_{\frac{(n+1)}{4}}$$

$$Q_3=x_{\frac{3(n+1)}{4}}$$

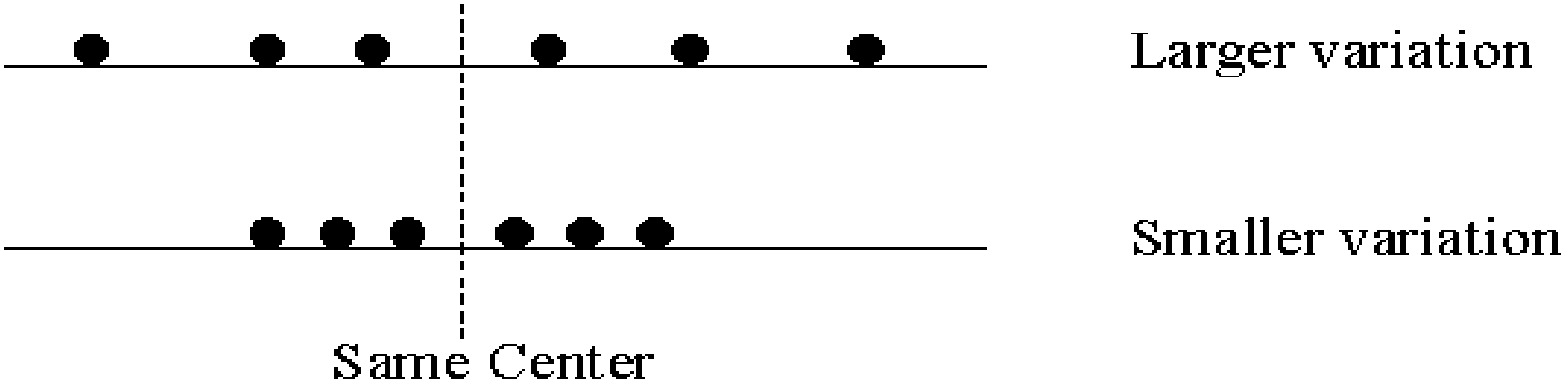
Measures of Dispersion

مقياس التشتت

- Mean and median measure tells useful information about central tendency, but they lack the ability to describe data variance. for example, data may have the same mean but totally different variance.
- Example 1 :
- Calculate mean for the following two sets of data .Set 1 : 100 10 50 50 90
- Set 2: 62 61 54 58 60
- Mean of set 1 = $\frac{100+10+50+90}{5} = \frac{300}{5} = 60$
- Mean of set 2 = $\frac{62+61+59+58+60}{5} = \frac{300}{5} = 60$

Measures of Dispersion

The variation or dispersion in a set of data refers to how spread out the observations are from each other.



The variation is small when the observations are close together. There is no variation if the observations are the same.

The Range:

- The difference between the largest and smallest sample values
- If $X_1, X_2, \ldots\ldots\ldots, X_n$ are the values of observations in a sample then range is given by:

$$Range(X_1, X_2, \ldots\ldots, X_n) =$$

$$\max(X_1, X_2, \ldots\ldots, X_n) - \min(X_1, X_2, \ldots\ldots, X_n)$$

The Range (Example):

find The range of (12, 24, 19, 20, 7) .

Solution:

$$Range = 24 - 7 = 17$$

- **One of the simplest measures of variability to calculate.**
- **Depends only on extreme values and provides no information about how the remaining data is distributed.**

range is simply the difference between the largest and smallest values in the dataset .

$$\textbf{Range}\ (\boldsymbol{x1}, \boldsymbol{x2}, \boldsymbol{x3}, \boldsymbol{x4}\ ) = \textbf{max}(\ \boldsymbol{x1}, \boldsymbol{x2}, \boldsymbol{x3}, \boldsymbol{x4}\ ) - \textbf{min}(\ \boldsymbol{x1}, \boldsymbol{x2}, \boldsymbol{x3}, \boldsymbol{x4}\ )$$

Example : find the ranges of the data in example 1

Set 1 : 100 10 50 50 90

Set 2: 62 61 59 58 60

Solation :

$$\min(100\ , 10\ , 50\ , 50\ , 90 = 10$$

$$\max(\ 100\ , 10\ , 50\ , 50\ , 90\ ) = 100$$

Set1 range = 100 – 10 = 90

$$\min(62\ , 61\ , 59\ , 58\ , 60 = 58$$

$$\max(\ 62\ , 61\ , 59\ , 58\ , 60\ ) = 62$$

Set2 range = 62 – 58 = 4

Set1 has high rang range which represents high variance and set 2 had low range (4) which represents low fluctuations.

Hint Range is a good measure for variability, but it is very weak if the data had outliers) (قيم شاذة the outliers affects the range but range in this case , will not express real variability.

القيم الشاذة تؤثر في المدى. على سبيل المثال إذا اردنا معرفة مدى الرواتب في مصنع من المصانع، وكان حوالي 100 عامل في المصنع يتقاضون رواتب بين 3 آلاف ريال وأربعة آلاف ريال. بذلك سيكون المدى بسيط وهو ألف ريال وهو مدى منطقي. أما إذا وجد مدير للمصنع يتقاضى 50 الف ريال. وهو الوحيد في المصنع الذي يتقاضى هذا الراتب. فإذا حسبنا المدى سيتأثربالقيمة الشاذة وهي قيمة الراتب 50 الف. وسيصبح 47 ألف ريال. وهو غير منطقي حيث أن التذبذب الخاص بالبيانات ليس مداه 50 ألف.

The following MATLAB command can be used to get the mode

$$\gg \mathbf{a} = [\mathbf{100} \quad \mathbf{10} \quad \mathbf{50} \quad \mathbf{50} \quad \mathbf{90}]$$

$$\gg \boldsymbol{c} = \mathbf{range}\ (\mathbf{a})$$

Variance and Standard Deviation

if (x1 , x2 , x3) is the data, then the variance an be calculated by

$$S^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2}{n - 1}$$

Where $\overline{x} = \frac{x_1 + x_2 + x_3}{2}$ which is the data average or mean. and *n* is the number of data values which is 3 in this case .

The sample variance is S^2

The population variance is σ^2

The variance is simply the average value of the distances between each value in the dataset and data-set central value. or the average value or the fluctuations variance is expressed mathematically as:

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

The variance is fine measure for describing variability, but the resulting value is squared, for example if the data set is heights of persons, the resulting values are not reflecting heights, it represent the square of the height.

Example: find the variance marks of the following sets which represents the degrees of STAT class students.

Set 1 : 100 10 50 50 90

Set 2 : 62 61 59 58 60

Set1 variance ($\bar{x} = 60$)

$$s_1^2 = \frac{1}{5-1}[(100-60)^2 + (10-60)^2 + (50-60)^2 + (50-60)^2 + (90-60)^2]$$

$$s_1^2 = 1300$$

Set2 variance ($\bar{x} = 60$)

$$s_1^2 = \frac{1}{5-1}[(62-60)^2 + (61-60)^2 + (59-60)^2 + (58-60)^2 + (60-60)^2]$$

$$s_2^1 = \frac{4+1+1+4+0}{4} = \frac{10}{4} = 2.5$$

So, set1 has higher variances (1300) them set2 (2.5)

Drawback: The problem is that 1300 is not a student mark, it doesn't reflect the same range of the data set , no one can get 1300 in the STAT Class !!

- To solve the above problem, the standard deviation is used, the standard divination is simply the square root of the variance .
- $s = \sqrt{s^2}$
- For set 1 $s = \sqrt{1300} = 36.056$
- For set 2 $s = \sqrt{2.5} = 1.5811$

- Now we can easily say that the deviation of date from the center is 36 marks .