Learning Outcome

By the end of this lecture, you will be able to:

- Compute and interpret the mean, median, mode, weighted mean, and percentiles for a set of data
- Compare different measures of location

Introduction

Let's consider a teacher's summary of a class's marks in mathematics. These marks give an idea about individual performance of every student.

Classroom

Student No.	Mathematics marks	Student performance	Class performance
1	34		?
2	56		
3	89		
4	99		
5	23		
6	67		
7	45		
8	78		
9	80		
10	56		

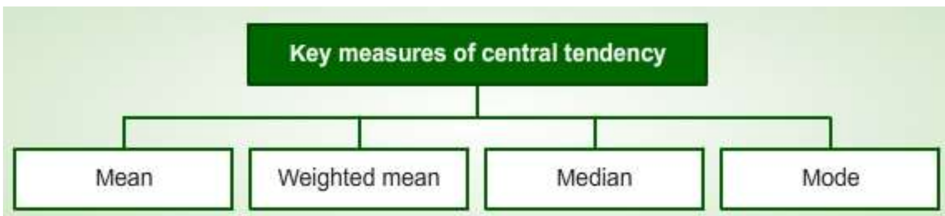
Marks in mathematics

What about performance of the entire class?

Well, the class teacher can use a statistical tool called measures of central tendency to assess the collective performance of the class. In this module, you will learn about the tools that help in the analysis of the distribution of given data. These include measures of location, measures of variation, and using some of the measures simultaneously.

Measures of Central Tendency

A measure of central tendency gives an indication of the point about which the data are gathered or clustered.



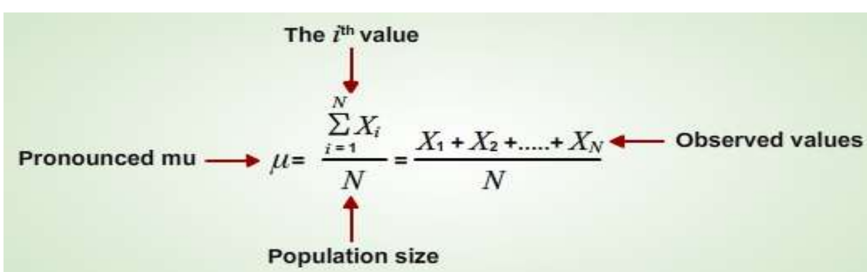
Key measures of central tendency

Mean > Definition and Types

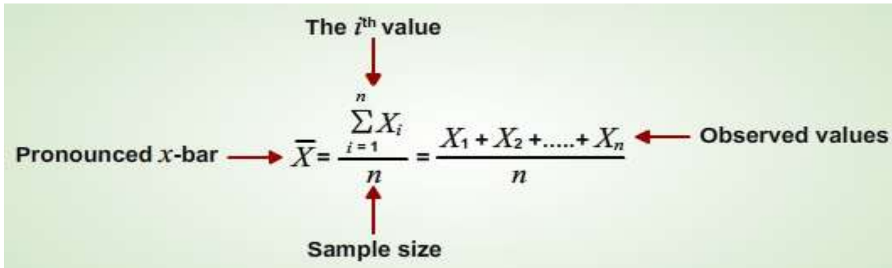
The mean or the arithmetic mean:

- The most common measure of central tendency
- The value which we expect on average and in the long run
- It equals the sum of values divided by the number of values

Finite population mean :



Sample mean:



Mean > Example

The following data shows the absences of all students in one section of Stat 211 course in a certain semester. Compute the population mean.

3	0	2	0	1	3	5	2
5	1	3	0	0	1	3	3
4	3	1	8	4	2	4	0

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\sum_{i=1}^{24} X_i}{24} = \frac{3 + 0 + \cdots + 0}{24} = \frac{58}{24} = 2.4167 \ absences$$

Mean > Example ...Contd.

The following are the average tips of twelve waiters on a usual working day in Saudi Riyals.

Waiter number	1	2	3	4	5	6	7	8	9	10	11	12
Tip (In Saudi Riyals)	10	13	12	12	13	15	12	12	13	15	2	20

Find the sample mean.

Solution:

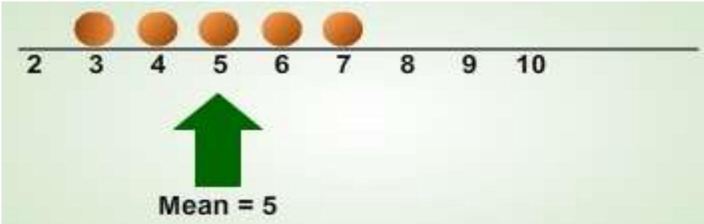
The mean is very sensitive to extreme values or outliers. Therefore, use other types of measures of central tendency, in case of extreme values or outliers in the data.

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{12} X_i}{12} = \frac{149}{12} = 12.417 \quad Saudi\ Riyals$$

Mean > Example ...Contd.

Find the mean for the set of data {3,4,5,6,7}. Explain whether it represents or does not represent the center of the data.

Solution:



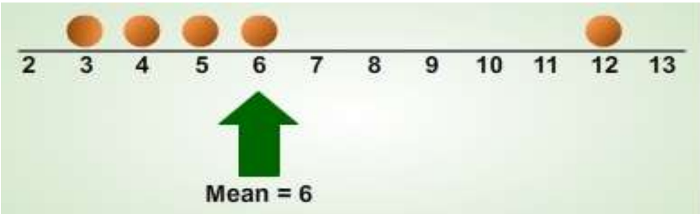
Example

Here, the mean is equal to 5, which is representative of the center of the data.

Mean > Example ...Contd.

Find the mean for the set of data {3,4,5,6,12} . Explain whether it represents or doesn't represent the center of the data.

Solution:



Mean

Here, the mean is equal to 6, which is not representative of the center of the data.

Weighted Mean

The weighted mean is used when values are grouped by frequency or relative importance. There is the weighted population mean and the weighted sample mean.

Population mean:

$$\mu_w = \frac{\sum w_i x_i}{\sum w_i}$$ w_i: the weight of i^{th} value.

Sample mean:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i}$$ w_i: the weight of i^{th} value

Weighted Mean > Example

A researcher surveyed twenty people to determine how many times they had meals outside the house per week in 2009. The results are recorded in the given frequency table. Find the average number times they had meals outside the house per week in 2009.

Frequency table

Number of meals	Frequency
0	6
1	4
2	6
3	3
4	1
Total	**20**

Solution:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum w_i} = \frac{0(6) + 1(4) + 2(6) + 3(3) + 4(1)}{20}$$
$$= \frac{29}{20} = 1.45 \ times\ per\ person$$

A group of sports-shoes stores recorded the profits from thirty different products as shown in the table.

Calculate mean of the given data.

Frequency table

Class	Frequency (in thousand dollars)
(10, 20]	5
(20, 30]	10
(30, 40]	6
(40, 50]	6
(50, 60]	3
Total	**30**

Sports-shoes

Solution:

Class	f	midpoint (x)	$x_i f_i$
(10, 20]	5	15	75
(20, 30]	10	25	250
(30, 40]	6	35	210
(40, 50]	6	45	270
(50, 60]	3	55	165
Total	**30**		970

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i f_i}{\sum f_i} = \frac{970}{30} = 32.333$$

in thousand dollars, which is an approximate mean. Note that, in case of grouped data, we use the mid-point of the interval and not the exact values, which are unknown, to calculate the approximate mean.

Median

The median of a sample of observations is the middle values.

Let X_1 , X_2 , ... , X_n represent a sample of size 'n'.
Then, we denote the ordered sample by $X_{(1)}$, $X_{(2)}$, ... , $X_{(n)}$;
Here

Minimum is denoted by $X_{(1)}$
Maximum is demoted by $X_{(n)}$

The middle value or median is denoted by 'm'.

	If n is odd	If n is even
Median (m)	$Middle\ value = X_{\left(\frac{n+1}{2}\right)}$	$average\ of\ the\ two\ middle\ values = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$

Note that the median is not affected by extreme values. It is computed from the center of the values. Further, median does not efficiently use information from all the data.

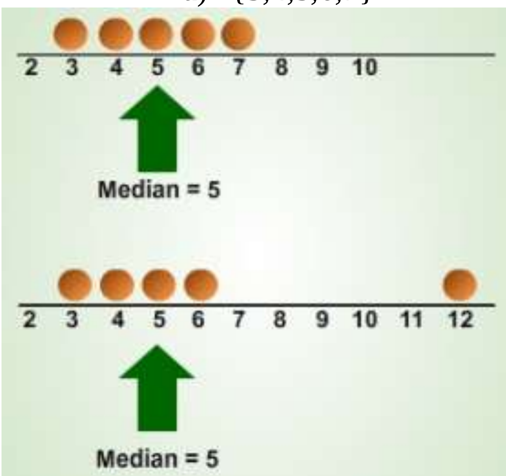
Median > Example

Find the median for the given sets of data.

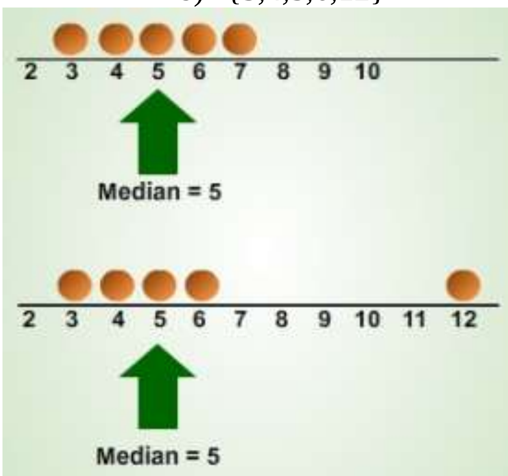
a) {3,4,5,6,7}
b) {3,4,5,6,12}

Solution:

a) {3,4,5,6,7}



b) {3,4,5,6,12}



For the first set of data, we can use the mean or the median. But, for the second list we use the median instead of the mean because of the extreme value.

Marks obtained by 7 students in STAT 211 are given by the table.

Student number	1	2	3	4	5	6	7
Marks	81	82	98	83	80	85	82

Find the median.

Solution:

1. Sort the observations:

   80 81 82 82 83 85 98

2. Since we have 7 observations (odd number), the location of the median is

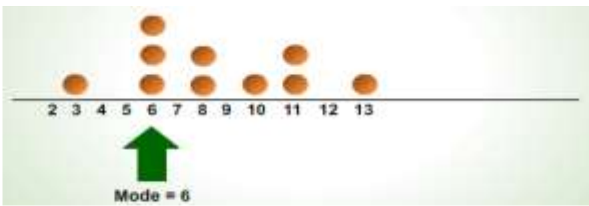
$$location\ of\ the\ median = \left(\frac{n+1}{2}\right)^{th} observation = 4^{th} observation$$

3. Hence, the median, $m = 82$

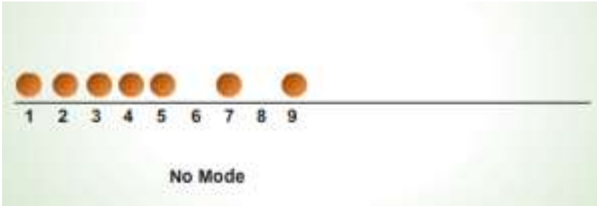
Mode

The mode is the most frequent value or the observation with the highest frequency. It is not affected by extreme values and is used for either numerical or categorical data. Note that there may be no mode at all or there may be several modes. For instance, observe the different mode values possible in the three cases.

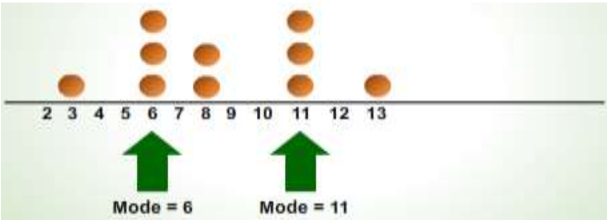
Single mode:



No mode:



Multiple modes:



Mode > Example

Assume that the grades for eight students are:

A+, B, A, D+, A ,B, D+, B+, B, B, A ,D+,F, D+, B, B, A+, D+, A.

Find the modal grade.

*Solution***:**

1. Find the frequency for each grade.

Grade	**Frequency**
A+	2
A	4
B+	1
B	6
D+	5
F	1

2. The largest frequency is 6
3. The mode = B

Comparison

The general comparison of the three measures of location is given.

1. The mean is generally used unless extreme values or outliers exist.
2. The median is often used since it is not sensitive to extreme values.
3. The mode is often used for categorical data.
4. In some situations, it makes sense to report both the mean and the median.

Comparison > Example

Consider the prices of five cars in a car show room:

Car prices

The average price is $6000
The median price is $3000
The modal price or the most frequent price is $1000

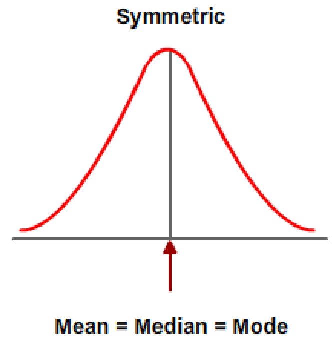
In this case, it is better to use the median or the mode instead of the mean as the mean is affected by extreme values.

Shape of a Distribution

The location of mean, median and mode values indicate the shape of a distribution. The shape of a distribution can be symmetrical, left-skewed, or right-skewed depending on location of mean, median and mode.

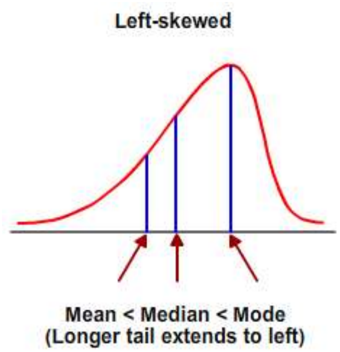
Symmetrical:

The data is spread, uniformly or regularly, around the center.



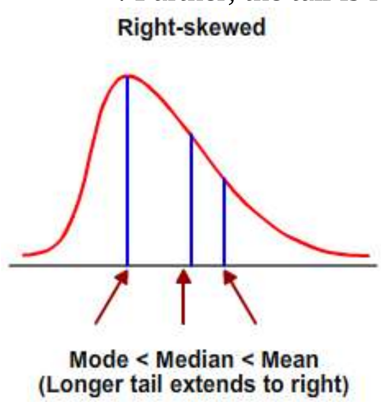
Left-skewed:

The data are not symmetric and $\bar{x} < m$. Further, the tail is longer and extended to left.



Right-skewed:

The data are not symmetric and $\bar{x} > m$. Further, the tail is longer and extended to right.



Percentiles

The α^{th} percentile (P_α) is the value that exceeds $\alpha\%$ of the data. It is obtained in four steps.

Step 1	Sort	Order the observations in an ascending order.
Step 2	Locate	Determine the location of the percentile. $L_\alpha = \alpha(n+1)/100 = i.d, \quad \alpha = 1, 2, \ldots, 99$.
Step 3	Separate	Separate i the largest integer not exceeding L_α and the decimal part(d) of L_α . If L_α is an integer, that is $d = 0$, then, the α^{th} percentile is the i^{th} observation. $P_\alpha = X_{(i)}$
Step 4	Calculate	Calculate the α^{th} percentile. $P_\alpha = x_{(i)} + d\,(x_{(i+1)} - x_{(i)}) = (1-d)x_{(i)} + dx_{(i+1)} \qquad \alpha = 1, 2, \ldots, 99.$

Quartiles

Quartiles are calculated by splitting a set of data into three equal parts. They are:

- **First Quartile:**

  1. P_{25}
  2. Denoted by Q_1
  3. Divides the smallest 25% of the values from the other 75% that the largest

- **Second Quartile:**

  1. P_{50}
  2. Denoted by Q_2
  3. Is the median

  50% of the values are smaller than the median and 50% are larger

- **Third Quartile:**

  1. P_{75}
  2. Denoted by Q_3
  3. Divides the smallest 75% of the values from the largest 25%

Quartiles > Example

The average tips of twelve waiters in a restaurant on a usual working day are given.

Waiter Number	1	2	3	4	5	6	7	8	9	10	11	12
Tip (in Saudi Riyals)	10	13	12	12	13	15	12	12	13	15	2	20

Find the 1st, 2nd, 3rd quartiles and the 90th percentile.

Solution:

1. Sort the data

$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$X_{(9)}$	$X_{(10)}$	$X_{(11)}$	$X_{(12)}$
2	10	12	12	12	12	13	13	13	15	15	20

2. First Quartile - $Q_1 = P_{25}$
   a. Find the location for Q_1
   b.

$$L_{25} = \frac{(n+1)\alpha}{100} = \frac{13(25)}{100} = 3.25$$

   c. $Q_1 = P_{25} = (1 - 0.25)X_{(3)} + 0.25\,X_{(4)} = 0.75(12) + 0.25(12) = 12$

3. Second Quartile - $Q_2 = P_{50}$
   a. Find the location for Q_2

$$L_{50} = \frac{(n+1)\alpha}{100} = \frac{13(50)}{100} = 6.5$$

   b. $Q_2 = P_{50} = (1 - 0.5)\,X_{(6)} + 0.5\,X_{(7)} = 0.5(12) + 0.5(13) = 12.5$

4. Third Quartile - $Q_3 = P_{75}$
   a. Find the location for Q_3

$$L_{75} = \frac{(n+1)\alpha}{100} = \frac{13(75)}{100} = 9.75$$

   b. $Q_3 = P_{75} = (1 - 0.75)X_{(9)} + 0.75\,X_{(10)} = 0.25(13) + 0.75(15) = 14.5$

5. 90th Percentile - P_{90}
   a. Find the location for P_{90}

$$L_{90} = \frac{(n+1)\alpha}{100} = \frac{13(90)}{100} = 11.70$$

   b. $P_{90} = (1 - 0.7)X_{(11)} + 0.7\,X_{(12)} = 0.3(15) + 0.7\,(20) = 18.5$

Recap

In this lecture, you have learned that:

- The key measures of central tendency are:
  - Mean
  - Weighted mean
  - Median
  - Mode

- The shape of a distribution denotes how data is distributed. It is symmetrical or left-skewed or right-skewed

- The α^{th} percentile (P_α) is the value that exceeds $\alpha\%$ of the data. It is obtained by following the four steps:
  - Sort
  - Locate
  - Separate
  - Calculate

- Quartiles are calculated by splitting a set of data into three equal parts:
  - First quartile
  - Second quartile
  - Third quartile