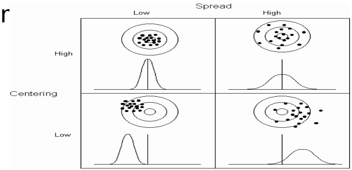
Applied Statistics I
(IMT224β/AMT224β)

Department of Mathematics
University of Ruhuna

A.W.L. Pubudu Thilan

Measure of central tendency, variation and shape

- You can characterize any set of data by measuring its central tendency, variation, and shape.
- Most popular measure of central tendency are the **mean, median** and **mode**.
- Variation measures the spread or dispersion of values in a data set.
- The **range, standard deviation** and **variation** are the commonly used in measure of variatior



The mean

- The **arithmetic mean** is the most common measure of central tendency.
- The mean serves as a balance point in a set of data.
- You can calculate the mean by adding together all the values in a data set and then dividing that sum by the number of values in the data set.
- If $x_1, x_2, ..., x_n$ are sample (or population) observations, then we can define mean as

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

Example

Suppose you define the time to get ready as the time in minute from when you get out of the bed to when you leave your home. You collect the times shown below for 10 consecutive works days. Calculate the mean time.

Day	1	2	3	4	5	6	7	8	9	10
Time	39	29	43	52	39	44	40	31	44	35

Cont...

$$\bar{x} = \frac{\text{Sum of the values}}{\text{Number of values}}$$
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$\bar{x} = \frac{39 + 29 + 43 + 52 + 39 + 44 + 40 + 31 + 44 + 35}{10}$$
$$\bar{x} = \frac{396}{10} = 39.6$$

The mean time is 39.6 minutes.

Arithmetic mean of grouped data

The mean when data are summarized with frequencies are given by

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}, \text{ k=no of values.}$$

Example 1

Consider the frequency distribution shown in scores of 20 students in a science test. Find the mean marks of the students?

Marks(x)	Frequency(f)
40	1
50	2
60	4
70	3
80	5
90	2
100	3
Total	20

Solution

Marks(x)	Frequency(f)	fx
40	1	40
50	2	100
60	4	240
70	3	210
80	5	400
90	2	180
100	3	300
Total	20	1470

$$\text{Mean marks} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{1470}{20} = 73.5$$

Mean for summarized data with classes

If we have summarized data with classes, we can use two methods to find mean.

- Direct method.
- Step deviation method.

Direct method

$$\text{Mean} = \frac{\sum_{i=1}^{k} f_i m_i}{\sum_{i=1}^{k} f_i};$$

Where
k=number of classes

f_i=frequency of i^{th} class

m_i=mid value of i^{th} class

Example

Compute the mean of the given data set.

Weight(Kg)	f_i
50.5− <53.5	1
53.5− <56.5	2
56.5− <59.5	6
59.5− <62.5	11
62.5− <65.5	16
65.5− <68.5	9
68.5− <71.5	4
71.5− <74.5	1

Solution

Weight(Kg)	f_i	m_i	$f_i m_i$
50.5− <53.5	1	52	52
53.5− <56.5	2	55	110
56.5− <59.5	6	58	348
59.5− <62.5	11	61	671
62.5− <65.5	16	64	1024
65.5− <68.5	9	67	603
68.5− <71.5	4	70	280
71.5− <74.5	1	73	73
	50		3161

Cont...

$$\bar{x} = \frac{\sum_{i=1}^{8} f_i x_i}{\sum_{i=1}^{8} f_i} = \frac{3163}{50} = 63.22$$

Step deviation method

$$
\begin{aligned}
\overline{x} &= A + \left(\frac{\sum_{i=1}^{k} f_i d_i}{\sum_{i=1}^{k} f_i} \right) w; \\
k &= \text{number of classes} \\
A &= \text{assumed mean} \\
w &= \text{width of the class intervals that it lies} \\
d_i &= \text{deviation of the } i^{th} \text{ class from the class that it lies.}
\end{aligned}
$$

Example

The emission of sulfur oxide of an industrial plant at 80 determinations are recorded and summarized by the following table.

Amount	f_i
5– <9	3
9– <13	10
13– <17	14
17– <21	25
21– <25	17
25– <29	9
29– <33	2

Find the mean.

Cont...

Amount	f_i	m_i	d_i	$f_i d_i$
5− <9	3	7	-3	-9
9− <13	10	11	-2	-20
13− <17	14	15	-1	-14
17− <21	25	**19**	0	0
21− <25	17	23	1	17
25− <29	9	27	2	18
29− <33	2	31	3	6
	80			-2

Cont...

Let $A = 19$

Let $17- < 21$be the class interval that A lies

$$\overline{x} = 19 + \left[\frac{(-2).4}{80}\right]$$
$$= 18.9$$

Properties of mean

- Mean always exists and it is unique.
- Mean depends on extreme values.
- It takes into account every item of data.

Weighted mean

- Arithmetic mean computed by considering relative importance of each items is called weighted arithmetic mean.
- Instead of each of the data points contributing equally to the final average, some data points contribute more than others.
- If all the weights are equal, then the weighted mean is the same as the arithmetic mean.

Weighted mean

If $x_1, x_2, ..., x_n$ are values, whose relative importance is expressed numerically by a corresponding set of numbers $w_1, w_2, ..., w_n$, then weighted mean $\overline{x}_w$, is given by

$$\overline{x}_w = \frac{x_1 w_1 + x_2 w_2 + ... + x_n w_n}{w_1 + w_2 + ... + w_n}.$$

Example

A student obtained $40, 50, 60, 80$, and 45 marks in the subjects of Math, Statistics, Physics, Chemistry and Biology respectively. Assuming weights $5, 2, 4, 3$, and 1 respectively for the above mentioned subjects. Find Weighted Arithmetic Mean per subject.

Solution

$$
\begin{aligned}
\overline{x}_w &= \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n} \\
&= \frac{40 \times 5 + 50 \times 2 + 60 \times 4 + 80 \times 3 + 45 \times 1}{5 + 2 + 4 + 3 + 1} \\
&= \frac{825}{15} \\
&= 55 \text{ marks/subject.}
\end{aligned}
$$

Median

- One type of average, found by arranging the values in order and then selecting the one in the middle.
- If the total number of values in the sample (or population) is even, then the median is the mean of the two middle numbers.
- The median is a useful number in cases where the distribution has very large extreme values which would otherwise skew the data.

Example 1

Find the median of the values $4, 1, 8, 13, 11$

Solution

Arrange the data $1, 4, 8, 11, 13$

$$\begin{aligned}\text{Median} &= \text{Value of } \left[\frac{n+1}{2}\right]^{th} \text{ item}\\ \text{Median} &= \text{Value of } \frac{6}{2} \text{ item} = 3^{rd} \text{ item}\\ \text{Median} &= 8\end{aligned}$$

Example 2

Find the median of the values $5, 7, 10, 20, 16, 12$.

Solution

Arrange the data $5, 7, 10, 12, 16, 20$

$$
\begin{aligned}
\text{Median} &= \text{Value of } \left[\frac{n+1}{2}\right]^{th} \text{ item} \\
\text{Median} &= \frac{7}{2}^{\text{th}} \text{ item} \\
&= 3.5^{\text{th}} \text{ item} \\
&= \frac{10+12}{2} = 11
\end{aligned}
$$

Example 3

Find the median of the following data set.

Weight (Kg)	No of students
20	5
22	7
23	4
24	1
27	3
28	4
30	1

Cont...

Weight (Kg)	No of students	Cumulative frequency
20	5	5
22	7	12
23	4	16
24	1	17
27	3	20
28	4	24
30	1	25

$$\begin{aligned} n &= 25 \\ \left(\frac{n+1}{2}\right) &= \left(\frac{26}{2}\right) = 13 \\ \text{Median} &= 23 \text{ (value of } 13^{th} \text{ item)} \end{aligned}$$

Median of grouped data

- The median for grouped data, we find the cumulative frequencies and then calculated the median number $\frac{n}{2}$.
- The median lies in the group (class) which corresponds to the cumulative frequency in which $\frac{n}{2}$ lies.

Cont...

- We use following formula to find the median,

$$\begin{aligned}
\text{Median} &= L_i + \left(\frac{n}{2} - c_{i-1}\right)\frac{w}{f_i}, \\
i &= \text{median class} \\
L_i &= \text{lower boundary of the median class} \\
w &= \text{class width of median class} \\
f_i &= \text{frequency of median class} \\
c_{i-1} &= \text{cumulative frequency of } (i-1)^{\text{th}} \text{ class.}
\end{aligned}$$

Example

Calculate median from the following data.

Group	Frequency
60-64	1
65-69	5
70-74	9
75-79	12
80-84	7
85-89	2

Cont...

Group	Frequency	Class boundary	Cumulative frequency
60-64	1	59.5 - 64.5	1
65-69	5	64.5 - 69.5	6
70-74	9	69.5 - 74.5	15
75-79	12	74.5 - 79.5	27
80-84	7	79.5 - 84.5	34
85-89	2	84.5 - 89.5	36

Cont...

$$
\begin{aligned}
\left(\frac{n}{2}\right)^{th} \text{ item} &= \frac{36}{2} = 18^{th} \text{ item} \\
\text{Median} &= L_i + \left(\frac{n}{2} - c_{i-1}\right)\frac{w}{f_i}. \\
&= 74.5 + \frac{5}{12}(18 - 15) \\
&= 74.5 + \frac{5}{12}(3) \\
&= 74.5 + 1.25 \\
&= 75.75
\end{aligned}
$$

Properties of median

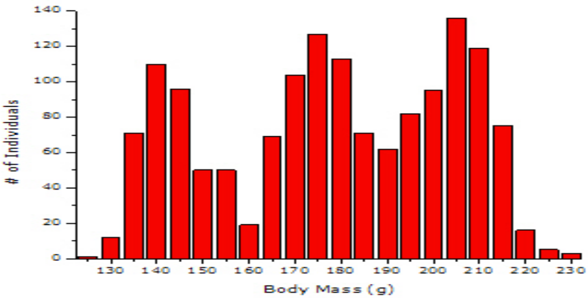
- It is unique and exists always.
- Has no effect from extreme values of the data set.
- It is not necessarily a particular observation of the data set.

Mode

- The mode is the value that occurs most frequently in a data set.
- There may be more than one mode when two or more numbers have an equal number of instances and this is also the maximum instances.
- A mode does not exist if no number has more than one instance.

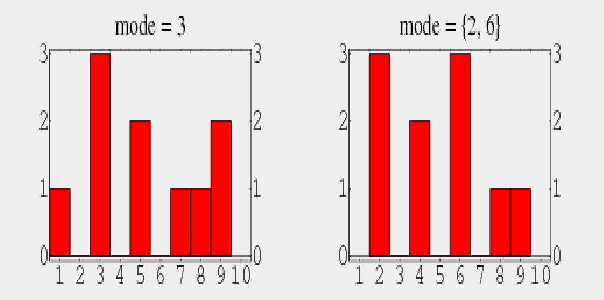
Cont...

- A distribution with a single mode is said to be unimodal. A distribution with more than one mode is said to be bimodal, trimodal, etc., or in general, multimodal.



Example 1

- For a data set, $3, 7, 3, 9, 9, 3, 5, 1, 8, 5$ the unique mode is 3 (left histogram).
- For a data set, $2, 4, 9, 6, 4, 6, 6, 2, 8, 2$ there are two modes: 2 and 6 (right histogram).



Example 2

Find the mode of the following data set.

Values	f_i
3	5
5	2
6	8
8	4
11	3

Note : For summarized data set with values, the mode is the most frequently occurring value.

Therefore mode is 6.

Mode for summarized data with class intervals

When the data are summarized with class intervals, the mode is given by

$$\begin{aligned} \text{Mode} &= L_i + \left[\frac{(f_i - f_{i-1})}{(f_i - f_{i-1}) + (f_i - f_{i+1})}\right] w \\ i &= \text{modal class} \\ L_i &= \text{lower boundary of modal class} \\ f_i &= \text{frequency of modal class} \\ f_{i-1} &= \text{frequency of } (i-1)^{th} \text{ class} \\ f_{i+1} &= \text{frequency of } (i+1)^{th} \text{ class} \\ w &= \text{class width of modal class.} \end{aligned}$$

Example

Calculate the mode of the given summarized data set.

Age	No of people
20− <25	60
25− <30	80
30− <35	100
35− <40	180
40− <45	150
45− <50	80
50− <55	120
55− <60	90

Cont...

Mode class is $35- < 40$.

$$\begin{aligned}
\text{Mode} &= L_i + \left[\frac{(f_i - f_{i-1})}{(f_i - f_{i-1}) + (f_i - f_{i+1})}\right] w \\
L_i &= 35 \\
f_i &= 180 \\
f_{i-1} &= 100 \\
f_{i+1} &= 150 \\
w &= 5. \\
\text{Mode} &= 35 + \left[\frac{(180 - 100)}{(180 - 100) + (180 - 150)}\right] 5 \\
&= 35 + \frac{80}{110} \times 5 \\
&= 38.635
\end{aligned}$$

Relationship

An interesting empirical relationship between the sample mean, statistical median, and mode which appears to hold for unimodal curves of moderate asymmetry is given by

$$\text{mode} \simeq \text{mean-3(mean-median)}.$$

Example

(a) For moderately skewed distribution mode=50.04, mean=45. Find median.

(b) If medain=20, and mean=22.5 in moderately skewed distribution then compute approximate value mode.

Solution

(a)

$$\begin{aligned} \text{mode} &\simeq \text{mean-3(mean-median)} \\ 50.04 &\simeq 45 - 3(45 - \text{median}) \\ \text{median} &\simeq 46.68 \end{aligned}$$

(b)

$$\begin{aligned} \text{mode} &\simeq \text{mean-3(mean-median)} \\ \text{mode} &\simeq 22.5 - 3(22.5 - 20) \\ \text{mode} &\simeq 15 \end{aligned}$$

Quartiles

- There are three quartiles called, first quartile, second quartile and third quartile.
- There quartiles divides the set of observations into four equal parts.
- The second quartile is equal to the median.
- The first quartile is also called lower quartile and is denoted by Q_1.

Cont...

- The third quartile is also called upper quartile and is denoted by Q_3.
- The lower quartile Q_1 is a point which has 25% observations less than it and 75% observations are above it.
- The upper quartile Q_3 is a point with 75% observations below it and 25% observations above it.

Quartile for ungrounded data

$$Q_1 = \text{Value of } \left[\frac{n+1}{4}\right]^{th} \text{ item}$$

$$Q_2 = \text{Value of } 2\left[\frac{n+1}{4}\right]^{th} \text{ item} = \text{Median}$$

$$Q_3 = \text{Value of } 3\left[\frac{n+1}{4}\right]^{th} \text{ item}$$

Example

The wheat production (in Kg) of 20 acres is given as:

1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730,
1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, 1885.

Find Q_1 and Q_3.

Solution

After arranging the observations in ascending order, we get

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$Q_1 = \text{Value of } \left[\frac{n+1}{4}\right]^{th} \text{ item}$$

$$Q_1 = \text{Value of } \left[\frac{20+1}{4}\right]^{th} \text{ item}$$

Cont...

$$
\begin{aligned}
Q_1 &= \text{Value of } [5.25]^{th} \text{ item} \\
Q_1 &= 5^{th} \text{ item} + 0.25(6^{th} \text{ item} - 5^{th} \text{ item}) \\
&= 1240 + 0.25(1320 - 1240) \\
&= 1240 + 20 \\
&= 1260
\end{aligned}
$$

Cont...

$$
\begin{aligned}
Q_3 &= \text{Value of } 3\left[\frac{n+1}{4}\right]^{th} \text{ item} \\
Q_3 &= \text{Value of } 3\left[\frac{20+1}{4}\right]^{th} \text{ item} \\
Q_3 &= \text{Value of } [15.75]^{th} \text{ item} \\
&= 15^{th} \text{ item} + 0.75(16^{th} \text{ item} - 15^{th} \text{ item}) \\
&= 1750 + 0.75(1755 - 1750) \\
&= 1753.75
\end{aligned}
$$

Example

The following table shows the distribution of 128 families according to the number of children.

No of children	No of families
0	20
1	15
2	25
3	30
4	18
5	10
6	6
7	3
8 or more	1

Find the quantile.

Cont...

No of children	No of families	Cumulative frequency
0	20	20
1	15	35
2	25	60
3	30	90
4	18	108
5	10	118
6	6	124
7	3	127
8 or more	1	128

Cont...

$$
\begin{aligned}
Q_1 &= \left(\frac{128+1}{4}\right)^{th} \text{ observation} \\
&= (32.25)^{th} \text{ observation} \\
&= 1 \\
Q_2 &= \left(\frac{128+1}{2}\right)^{th} \text{ observation} \\
&= (64.5)^{th} \text{ observation} \\
&= 3 \\
Q_3 &= 3\left(\frac{128+1}{4}\right)^{th} \text{ observation} \\
&= (96.75)^{th} \text{ observation} \\
&= 4
\end{aligned}
$$

Quantile for summarized data with class intervals

$$Q_1 = L_i + \left(\frac{n}{4} - c_{i-1}\right)\frac{w}{f_i};$$

$$\begin{aligned} L_i &= \text{lower boundary of the class in which } Q_1 \text{ lies} \\ f_i &= \text{frequency of that class} \\ w &= \text{width of that class} \\ c_{i-1} &= \text{cumulative frequency of proceeding class.} \end{aligned}$$

Cont...

$$Q_2 = L_i + \left(\frac{n}{2} - c_{i-1}\right)\frac{w}{f_i};$$

$$\begin{aligned}
L_i &= \text{lower boundary of the class in which } Q_2 \text{ lies} \\
f_i &= \text{frequency of that class} \\
w &= \text{width of that class} \\
c_{i-1} &= \text{cumulative frequency of proceeding class.}
\end{aligned}$$

Cont...

$$Q_3 = L_i + \left(\frac{3n}{4} - c_{i-1}\right)\frac{w}{f_i};$$

$$\begin{aligned} L_i &= \text{lower boundary of the class in which } Q_3 \text{ lies} \\ f_i &= \text{frequency of that class} \\ w &= \text{width of that class} \\ c_{i-1} &= \text{cumulative frequency of proceeding class.} \end{aligned}$$

Example

Calculate the quartile from the data given below:

Maximum Load	Number of Cables
9.3-9.7	2
9.8-10.2	5
10.3-10.7	12
10.8-11.2	17
11.3-11.7	14
11.8-12.2	6
12.3-12.7	3
12.8-13.2	1

Cont...

Maximum Load	No of Cables	Class boundary	C.F
9.3-9.7	2	9.25-9.75	2
9.8-10.2	5	9.75-10.25	7
10.3-10.7	12	10.25-10.75	19
10.8-11.2	17	10.75-11.25	36
11.3-11.7	14	11.25-11.75	50
11.8-12.2	6	11.75-12.25	56
12.3-12.7	3	12.25-12.75	59
12.8-13.2	1	12.75-13.25	60

Cont...

$$
\begin{aligned}
Q_1 &= \text{value of } \left(\frac{n}{4}\right)^{th} \text{ item} \\
Q_1 &= \text{value of } \left(\frac{60}{4}\right)^{th} \text{ item} \\
&= 15^{th} \text{ item} \\
Q_1 &\Rightarrow \text{ lies in the class } 10.25 - 10.75 \\
Q_1 &= 10.25 + (15 - 7)\frac{0.5}{12} \\
&= 10.25 + 0.33 = 10.58
\end{aligned}
$$

Cont...

$$
\begin{aligned}
Q_3 &= \text{value of } \left(\frac{3n}{4}\right)^{th} \text{ item} \\
Q_1 &= \text{value of } \left(\frac{3 \times 60}{4}\right)^{th} \text{ item} \\
&= 45^{th} \text{ item} \\
Q_3 &\Rightarrow \text{ lies in the class } 11.25 - 11.75 \\
Q_3 &= 11.25 + (45 - 36)\,\frac{0.5}{14} \\
&= 11.25 + 0.32 = 11.57
\end{aligned}
$$

Percentile

- A percentile is the value of a variable below which a certain percent of observations fall.
- For example, the 20^{th} percentile is the value (or score) below which 20 percent of the observations may be found.

Example

Consider the marks of students for MCQ paper of 40 questions. The marks are recorded as followings.

Marks	Number of Students
0-5	3
6-10	10
11-15	14
16-20	20
21-25	13
26-30	9
31-35	1

Find quartiles and 45^{th} percentile.

Solution

In finding percentiles we use the graph of percentage cumulative frequency polygon.

Marks	Number of Students	CF
0-5	3	3
6-10	10	13
11-15	14	27
16-20	20	47
21-25	13	60
26-30	9	69
31-35	1	70

Cont...

Left class boundaries	CF	CF%
0	0	0.00
5.5	3	4.28
10.5	13	18.57
15.5	27	38.57
20.5	47	67.14
25.5	60	85.71
30.5	69	98.57
35.5	70	100.00

Thanks