
STATISTICS

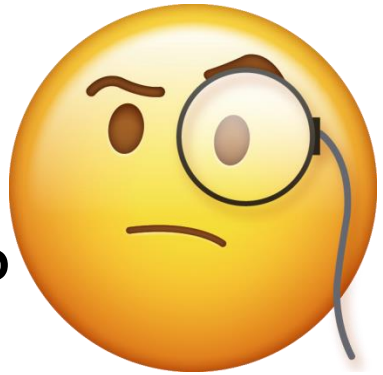
Measures of central location.

They are generally referred to as averages.

Used to describe the center of a data set.

In our daily lives…we often encounter expressions like…The average daily temperature…the average monthly rainfall…and the average weekly wage…

…but what do these actually mean…?



Suppose for instance…that the average weekly wage at a certain factory is ¥500…What does this then mean?

1. This is the middle wage. (half the employees earn less than 500 while the other half earn more than 500?)
2. This is the most common wage (more employees earn this wage than any other amount)
3. This is the wage lying mid way between the lowest and highest wages
4. Something else perhaps?

So the word average can be very misleading. We need to carefully define how we use this word.

We will consider the following…

Mean…Median & Mode.

These measures are precise…and easy to manipulate.


THE AVERAGE

The mean is better known as the average.

An easy way to think about the mean…is…

Imagine all of the money in everyone's pocket…being put into a large pile…and then that pile of money is shared evenly.

This share is then known as the average.

Thus the process is therefore…add and then divide.

The mean can be calculated as follows:

$$mean = \frac{Sum\ of\ values}{Number\ of\ values}$$

We simply add the data values and divide by the number of values.

The mean is the most common measure of the center of a set of data.

If we then calculate the mean using sample values…then the mean is referred to the sample mean.

Suppose we have a sample data set of:

$$x_1 \ldots x_2 \ldots x_3 \ldots x_n$$

$$\overline{x}$$

$$\overline{x} = \frac{Sx}{n}$$

Find the mean of:

A) 5, 6, 7, 9, 11, 12, 15

B) 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 6, 7, 8, 8

Find the mean of:

A) 5, 6, 7, 9 11, 12, 15

$$\overline{x} = \frac{5+6+7+9+11+12+15}{7} = \frac{65}{7} \approx 9.29$$

B) 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 6, 7, 8, 8

$$\overline{x} = \frac{6\times3+3\times4+2\times5+5\times6+7+2\times8}{19}$$

$$\approx 4.89$$

Over a 7 day period…the number of customers using a certain café was…

92…84…70…76…66…80…and 71

Calculate the sample mean

$$\bar{x} = \frac{Sx}{n}$$

$$= \frac{92 + 84 + 70 + 76 + 66 + 80 + 71}{7}$$

$$= \frac{539}{7}$$

$$= 77$$

So this means that the average number of customers is 77 per day.

The mean has a number of interesting properties…

The difference $\boldsymbol{x - \overline{x}}$ between any data value x and the mean $\overline{x}$ is called the deviation of that value away from the mean.

The deviations always sum to zero.

$$\boldsymbol{S\,(x - \overline{x}) = 0}$$

In our previous example… $\overline{x} = 77$

The 7 deviations of the 7 data values from the mean are:

$$92 - 77 = 15$$
$$76 - 77 = -1$$
$$71 - 77 = -6$$
$$84 - 77 = 7$$
$$66 - 77 = -11$$
$$70 - 77 = -7$$
$$80 - 77 = 3$$

The sum of these deviations are:

$$15 + 7 - 7 - 1 - 11 + 3 - 6 = 0$$

The negative deviations cancel the positive deviations.

In statistics, we will often be interested in the mean of a population.

It is defined in the same way, but different symbols are used. The mean of a population of N values is given by the formula:

$$\mu = \frac{SX}{N}$$

Where μ is the lower Greek letter mu (pronounced mew)

The mean is simple to calculate and interpret.

It uses all of the data values and can be calculated exactly.

However it is affected by outliers.

Extreme values…whether large or small can greatly distort the center value.

Task:

1. **Find the mean of:**

a) 5…6…7…9…11…12…15

b) 20…22…24…25…27…31…33

c) 14.5…17.5…18.5…21.5…22.5…23.5…25.5…26.5…

d) 150…154…158…160…166…170…173…

2. Find the mean of:

a) 30cm…45cm…15cm…25cm…36cm…

b) 24°C…22°C…28°C…23°C…21°C…19°C…

c) 120kg…115kg…125kg…130kg…132kg…138kg …148kg…110kg…126kg…

3. 6 different cans of cat food were purchased as part of a consumer price survey. The cans cost:

78c…106c…98c…92c…85c…88c…

Calculate the mean price per can.

4. Calculate the mean of
3…3…3…3…3…3…4…4…4…5…5…6…6…6…6…6…7…8…8…

5. Find the sum of the deviations of the data values 150…154…158…160…166..170…173…


THE MEDIAN

In this section…we will discuss the median.

When the data is arranged in ascending or descending order…the median is the middle score.

The median which can be denoted as M_d, partitions the data into 2 equal portions

When the median and the mean are very different…we call this **SKEWED.**

When the median and mean are very similar…we call this **SYMMETRICAL.**

When there is an odd number of data values…the median is equal to the middle value.

When there is an even number of data values, the median is defined as the value mid-way between the 2 middle values.

In both cases…the median of a set of n data values can be found using the single formula:

$$\boldsymbol{M_d = x_{(n+1)/2}}$$

The subscript indicates that the median is $\frac{1}{2}(n+1)$ data value.

Find the median of the following data set.

66…71…80…84…92…70…76…

First arrange the data is numerical order…

66…70…71…76…80…84…92…

Since n = 7 which is an odd number…

The median is $\frac{1}{2}(7+1) = 4^{th} data\ value$

…and the 4^{th} data value is 76

Notes:

The median 76 is slightly smaller than the mean…which is 77.

Thus the mean and median are not necessarily equal.

When the formula is used…the subscript 4 tells us that the median is the 4^{th} data value.

IT DOES NOT TELL US THAT THE MEDIAN IS 4.

Find the median of the numbers

5…11…8…7…9…4

First arrange the number in numerical order…

4…5…7…8…9…11…

Since n = 6…which is an even number…the median is the $\frac{1}{2}(6+1) = 3.5^{th} data\ value$

Thus…

$$M_d = \frac{7+8}{2}$$
$$= 7.5$$

When an even number of data values is involved…do you agree that any number between the 2 middle values would satisfy the requirement that half the values lie on either side of the median?

The median is simple to understand.

It is only concerned with the middle value.

It is unaffected by extremely large or small values.

However it cannot be used easily in other statistical calculations.

1. Determine the location of the median of a set of:

a) 13 data values

b) 24 data values

2. Find the median of:

a) 2…0…5…8…4…6…9…

b) 12.2…11.4…15.6…15.6…12.5…11.9…16.8…14.1…

3. Find the median of:

a) 1…7…4…15…11…19…

b) 1…7…4…15…11

4. The heights of 30 students, measured to the nearest centimeter are:

171	166	162	170	172	164	174	189	175	168
174	177	175	166	178	176	182	168	183	173
168	180	191	193	169	196	170	181	178	177

Find the median Height.

5. Find the mean and the median of:

2…3…4…6…19…3…5…5…8…4…4…

And discuss why the median is a more appropriate average than the mean.


THE MODE

A third common measure of central location is the mode…which can be denoted by M_o.

The mode is the most typical or most popular value of a set of data.

It is the value that occurs most frequently

It requires no calculation

A set of data is said to be **UNIMODAL** if there is **ONE MODE**…**BIMODAL** is there are two values with the same highest frequency etc.

Find the mode of each of the following data sets:

a) 3…5…1…2…7…5…6…3..5…4…5…5…
b) 1…7…15…9…3…4…5…8…
c) 2…2…3…4…6…2…4…5…5…4…

a) The mode is 5. This is the value that occurs most often.

b) The mode here does not exist. All values occur with equal frequency.

c) There are two modes…2 and 4.

The mode is also simple to understand and is not affected by extreme values.

The mode is unsuitable for further calculations.

The mode does not necessarily lie at the center of a set of data…although it does indicate where the greatest cluster of values is located.

1. Find the mode of:

a) 7…8…5…1…6…3…5…4…2…

b) 21…24…27…25…26…24…25…

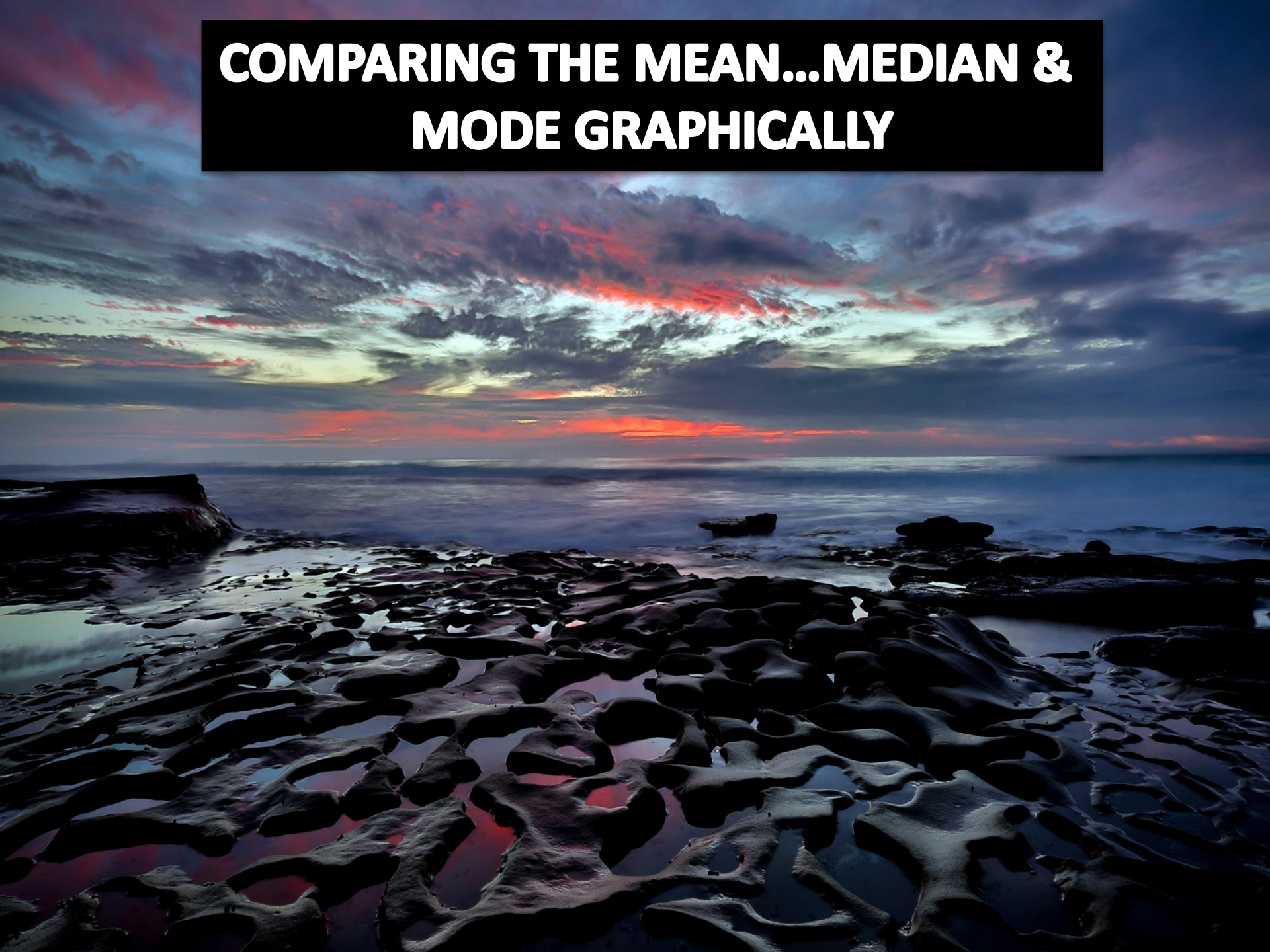
c) 105…100…110…110…105…120…115…115…110…125…

d) 11…14…18…21…27…34…

2. Find the mode of:

1…1…2…2…2…3…3…4…4…4…5…5…5…5…5…6…6…7…

COMPARING THE MEAN…MEDIAN & MODE GRAPHICALLY



We can use the mean…median and mode of a set of data and the relationships between their values to provide some information about the shape of the sample (or population)

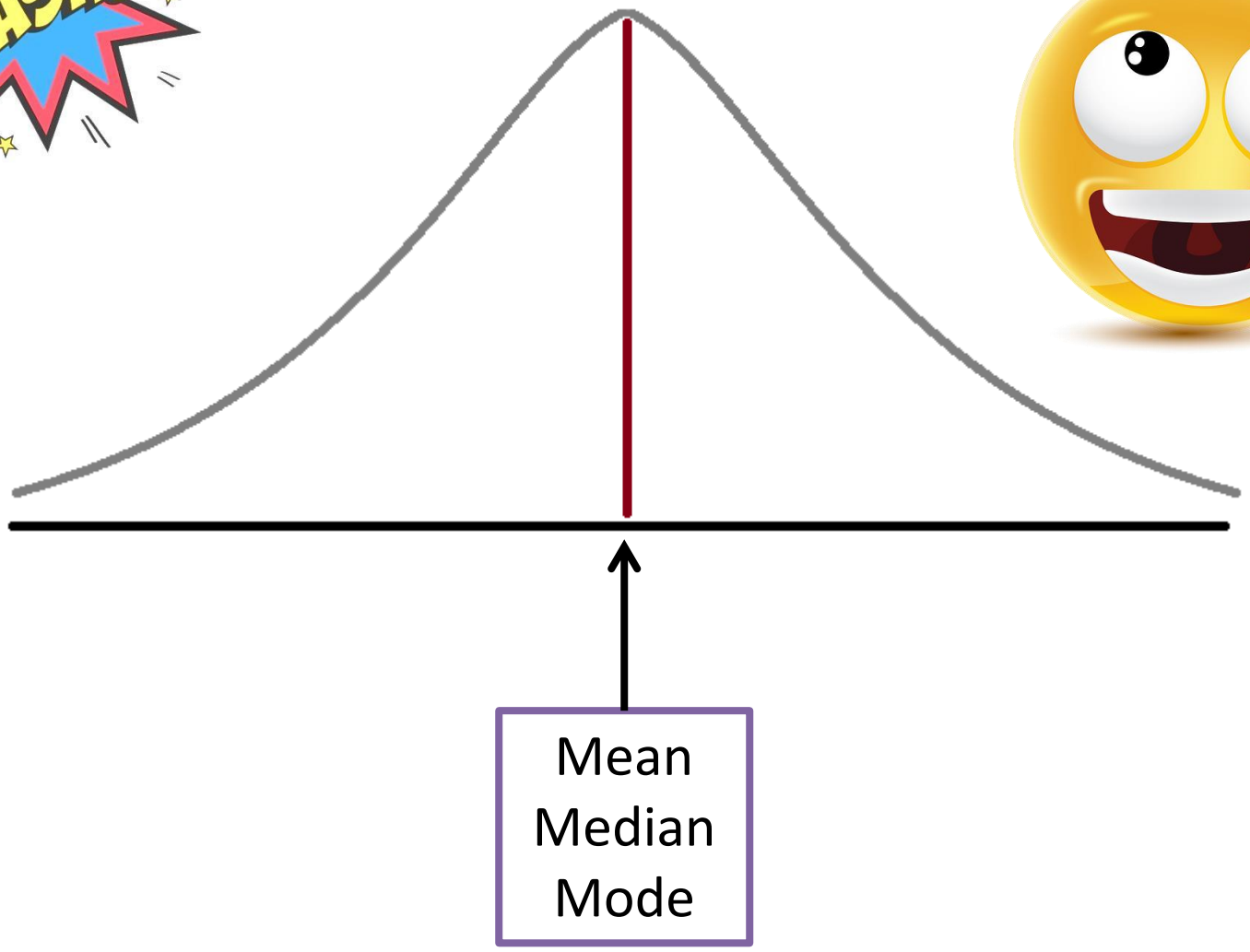
For a symmetrical bell shaped distribution, the mean…median and mode are equal.

All are located under the peak of the curve…

Thus…

$$\boldsymbol{\mu = M_d = M_o}$$


CRASH!


Mean
Median
Mode

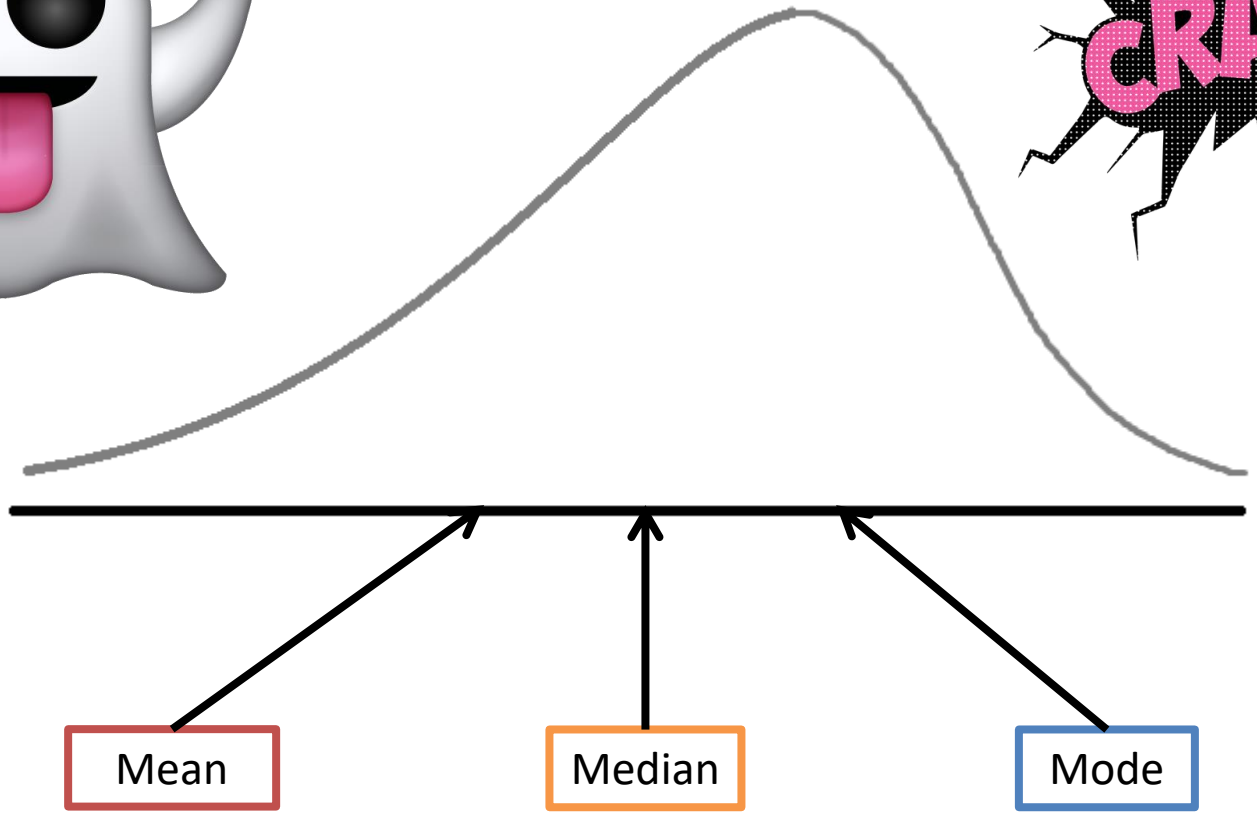
For a distribution that is skewed to the left….the mean is smaller than the mode since it is affected by the smaller values in the tail on the left.

The mode is located under the peak and the median is somewhere in between.

Thus for a distribution that is skewed to the left:

$$\boldsymbol{\mu < M_d < M_o}$$


CRACK!

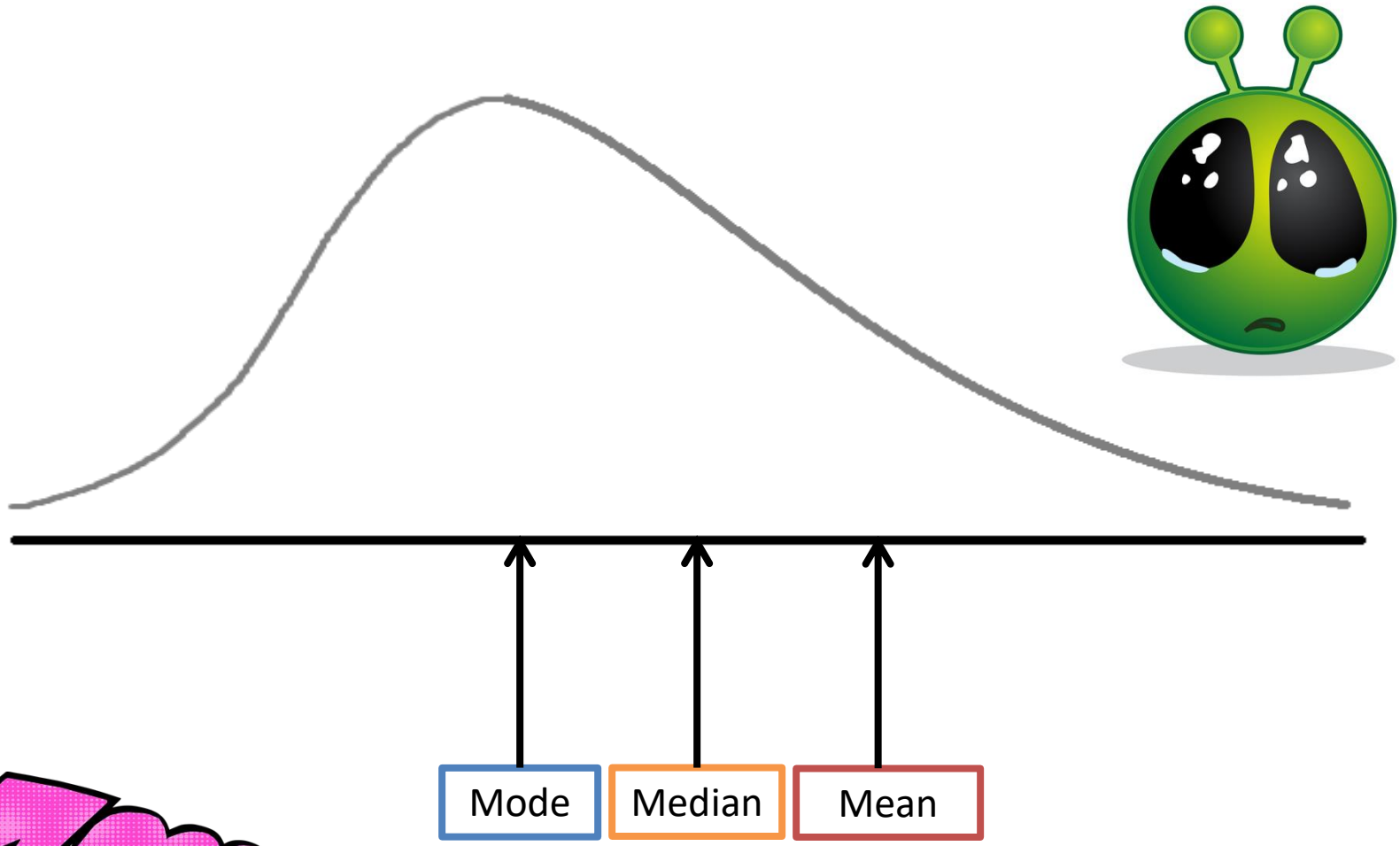

Mean
Median
Mode

For a distribution that is skewed to the right…the mean is larger than the mode since it is affected by the larger values in the tail on the right.

The mode is located under the peak and the median is somewhere in between.

Thus for a distribution skewed to the right…

$$M_o < M_d < \mu$$


Mode
Median
Mean


Zoooom!!

In each of the cases…observe that the mode is always at the peak (greatest frequency) and that the median always lies between the mean and the mode.

PEARSONS COEFFICIENT OF SKEWNESS allows us to numerically represent the degree to which a data set is skewed. It makes use of the mean…median (or mode) and one other measure called **STANDARD DEVIATION.**

$$\boldsymbol{Pearsons\ coefficient\ of\ Skewness} = \frac{\boldsymbol{mean\ - mode}}{\boldsymbol{Standard\ deviation}}$$

$$= \frac{\boldsymbol{3(mean\ - median)}}{\boldsymbol{Standard\ deviation}}$$

From this, we can readily see that a distribution that is skewed to the right (positively skewed) has a positive value for Pearson's Coefficient.

Similarly, a distribution that is skewed to the left (negatively skewed) has a negative value for Pearson's Coefficient.

In certain circumstances, it is important to know which of the 3 measures of central tendency **BEST REPRESENTS** a set of data.

This depends on the purpose to which the measure is to be put.

As discussed already…each measure has a number of advantages and disadvantages.

The mean is the best measure to use for the purpose of statistical inference.

However the median and mode are more useful measure for some purposes.

For descriptive purposes, it is usually best to report the values of all 3 measure since each conveys slightly different information.

1. Discuss the shape of the distribution with…

a) Mean = 100…median = 100…mode = 100

b) Mean = 24.3…median = 28.1…mode = 32.5

c) Mean = 56…median = 52…mode = 44

2. Jeremiah scored 25…30..15…28…and 75 marks in 5 tests.

Find the mean…median and mode of his test marks.

Which is the best average to use?

3. Jessica scores were 46…60…55…60…and 70.

Find the mean…median and mode of her test marks.

Which is the best average to use?

This section is about quartiles.

The first quartile is found like this:

Add 1 to the number of numbers, divide this by 4 and find the number at this position.

$$Q_1 = X_n + 1/4$$

There are 25% of the numbers below this value.

For the 3rd quartile:

Add 1 to the number of numbers, multiply by 3 and divide by 4, and find the number at this position.

$$Q_1 = 3_{X_n+1}/4$$

There are 75% of numbers below this value.

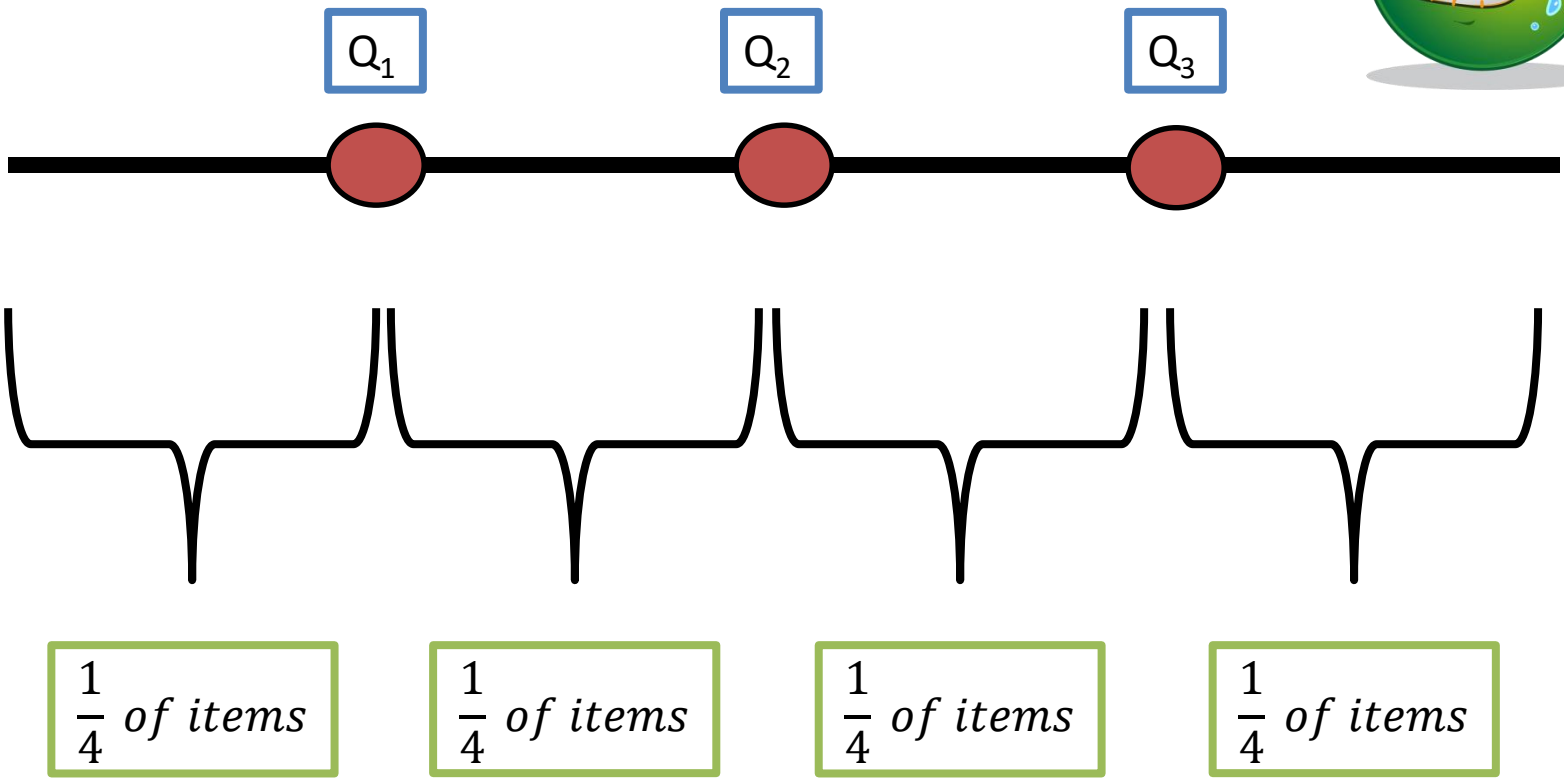
Measure of central location are single numbers used to describe the location of the Center of a set of data. It is often helpful to use more than one number to describe the location of particular data values in relation to the other values.

For example…suppose you believer that you are very much underpaid compared to other people with the same qualifications, experience and performance. By obtaining salaries paid to the other employees, you would be able to compare your salary to that of the entire group by using a measure of position to show that you are well down the range of salaries.

We have seen that the median divides the set of data into 2 equal parts…

We can extend this idea by dividing the data set into as many parts as we want.

Quartiles are the 3 values that partition a set of data values into 4 equal parts…each containing 25% of the values.



- One quartile or 25% of the data values lie below the first or lower quartile…denoted by Q_1.
- The second…denoted by Q_2…is the median…M_d.
- Half…or 50% of the data values lie below and above Q_2.
- The third or upper quartile denoted by Q_3…has 75% of the data values below it.

We can use the same procedure to find quartiles as we do to find the median.

If a set of n data values is arranged in ascending order and partitioned into 4 equal parts, then:

$$\boldsymbol{Q_1\ is\ the\ \frac{1}{4}(n+1)^{th} value\ \ldots so\ Q_1 = x_{(n+1)/4}}$$

$$\boldsymbol{Q_3\ is\ the\ \frac{3}{4}(n+1)^{th} value\ \ldots so\ Q_3 = x_{3(n+1)/4}}$$

From this pattern we can see that…

$$Q_2 = x_{2(n+1)/4} = x_{(n+1)/2} = M_d$$

Notes:

Quartiles will be used in measuring the variability or the spread of data.

Quartiles can also be used to find the degree of skewness in a dataset.

$$\boldsymbol{Quartile\ measure\ of\ skewness} = \frac{\boldsymbol{Q_3 - 2Q_2 + Q_1}}{\boldsymbol{Q_3 - Q_1}}$$

Find the 1st and 3rd quartiles of the data values:

5…8…12…17…21…22…24…24…28…30…

Since n = 10 and the data values are already arranged in ascending order…

$$Q_1 is\ the\ \frac{1}{4}(10+1)^{th} or\ 2.75^{th} data\ value\ \ldots where:$$

$$\begin{aligned} Q_1 &= X_{(n+1)/4} \\ &= X_{11/4} \\ &= X_{2.75} \end{aligned}$$

$$= x_2 + \frac{3}{4}(x_3 - x_2)$$

$$= 8 + \frac{3}{4}(12 - 8)$$

$$= 8 + 3$$

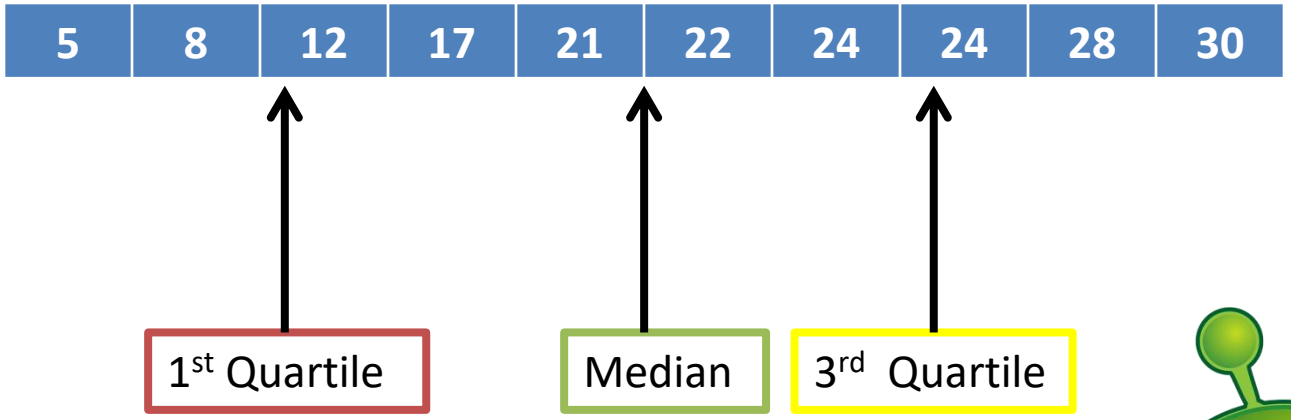
$$= 11$$

$Q_3 is\ the\ \frac{3}{4}(10+1)^{th} or\ the\ 8.25^{th} data\ value\ \ldots$

$$Q_3 = x_{3(n+1)/4}$$
$$= x_{33/4}$$
$$= x_{8.25}$$

$$= x_8 + \frac{1}{4}(x_9 - x_8)$$
$$= 24 + \frac{1}{4}(28 - 24)$$
$$24 + 1$$
$$= 25$$

As with the median…we can illustrate the position of the quartiles in relation to the data values as illustrated in the following diagram…



1. Find the 1st quartile Q_1 and the third quartile Q_3 for:

a) 4…7…9…12…16…18…18…18…

b) 20…22…23…24…29…29…

2. A certain small business has 10 employees.

Last month these employees were absent from work for a number of days because of illness.

Calculate Q_1 and Q_3 using the following data:

6	8	2	1	12	3	11	4	5	7

FIVE NUMBER SUMMARY

As well as the values at 25%...50% and 75%...we include the lowest (0%) and the highest (100%) to give the 5 number summary.

We often use a 5 number summary to describe or summarize a set of data.

The summary consists of the 5 statistics below arranged in order from the smallest to the largest:

1. Lowest value…x_L
2. 1st Quartile…Q_1
3. Median…M_d
4. 3rd Quartile…Q_3
5. Highest value…x_H

1. Find the five number summary of:

15…23…26…16…30…61…28…35…14…8

2. The monthly rent in dollars of flats in the locality of a certain university are:

540…545…555…560…560…570…575…590…650…730…

Find the 5 number summary

ANSWER SECTION

1. a)

$$Mean = \frac{S_x}{n}$$
$$= \frac{5 + 6 + 7 + 9 + 11 + 12 + 15}{7}$$
$$= 9.2857 = 9.3$$

b)

$$Mean = \frac{S_x}{n}$$
$$= \frac{20 + 22 + 24 + 25 + 27 + 31 + 33}{7}$$
$$= 26$$

c)

$$Mean = \frac{S_x}{n}$$
$$= \frac{14.5 + 17.5 + 18.5 + 21.5 + 22.5 + 23.5 + 25.5 + 26.5}{8}$$
$$= 21.25$$

d)

$$Mean = \frac{S_x}{n}$$
$$= \frac{150 + 154 + 158 + 160 + 166 + 170 + 173}{7}$$
$$= 161.6$$

2. a)

$$Mean = \frac{S_x}{n}$$

$$= \frac{30 + 45 + 15 + 25 + 36}{5}$$

$$= 30.2cm$$

b)

$$Mean = \frac{S_x}{n}$$

$$= \frac{24 + 22 + 28 + 23 + 21 + 19}{6}$$

$$= 22.833 = 22.8^oC$$

c)

$$Mean = \frac{S_x}{n}$$

$$= \frac{120 + 115 + 125 + 130 + 132 + 138 + 148 + 110 + 126}{9}$$

$$= 127.1kg$$

d)

$$Mean = \frac{S_x}{n}$$

$$= \frac{1300 + 1325 + 1400 + 1420 + 1475 + 1510 + 1555 + 1600 + 1660 + 1705 + 2000}{11}$$

$$= 1541$$

3.

$$Mean = \frac{S_x}{n}$$

$$= \frac{78 + 106 + 98 + 92 + 85 + 88}{6}$$

$$= 91.167$$

4.

$$Mean = \frac{S_x}{n}$$

$$= \frac{(3 \times 6) + (4 \times 3) + (5 \times 2) + (6 \times 5) + (7 \times 1) + (8 \times 2)}{19}$$

$$= 4.89$$

5.

$$Mean = \frac{S_x}{n}$$

$$= \frac{150 + 154 + 158 + 160 + 166 + 170 + 173}{7}$$

$$= 161.571$$

$$S(x - Mean)$$

$$= (150 - 161.571) + (154 - 161.571)$$
$$+ (158 - 161.571) + (160 - 161.571)$$
$$+ (166 - 161.571) + (170 - 161.571)$$
$$+ (173 - 161.571)$$

$$= 0$$

1. a)

$$n = 13 \ldots so\ median\ is \ldots$$

$$M_d = x_{(n+1)/2}$$

$$= x_{(13+1)/2}$$

$$= x_7$$

Which is the 7th data value.

b)

$$n = 24 \ldots so\ median\ is \ldots$$

$$M_d = x_{(n+1)/2}$$

$$= x_{(24+1)/2}$$

$$= x_{12.5}$$

Which is the 12.5th data value. Midway between 12th and 13th data values.

2. a)

First arrange the n = 7 data values in order of magnitude as:

0…2…4…5…6…8…9

$$M_d = x_{(n+1)/2}$$
$$= x_{(7+1)/2}$$
$$= x_4$$

The 4th data value is 5…thus the median is 5!

b)

The n = 8 data values in order of magnitude are:

11.4…11.9…12.2…12.5…14.1…15.6…15.6…16.8…

$$M_d = x_{(n+1)/2}$$
$$= x_{(8+1)/2}$$
$$= x_{4.5}$$

Answer = 13.3

3. a)

The n = 6 data values are:

1…4…7…11…15…19…

$$M_d = x_{(n+1)/2}$$
$$= x_{(6+1)/2}$$
$$= x_{3.5}$$

Thus the answer is 9

b) The n = 5 data values are:

1…4…7…11…15…

$$M_d = x_{(n+1)/2}$$
$$= x_{(5+1)/2}$$
$$= x_3$$

The answer is 7

The n = 30 values in order of magnitude are:

162	164	166	166	168	168	168	169	170	170
171	172	173	174	174	175	175	176	177	177
178	178	180	181	182	183	189	191	193	196

$$M_d = x_{(n+1)/2}$$
$$= x_{(30+1)/2}$$
$$= x_{15.5}$$

Thus the answer is 174.5cm

5. Arrange the n = 11 values in order of magnitude as:
2…3…3…4…4…4…5…5…6…8…19…

$$mean = \frac{S_x}{n}$$

$$= \frac{(2 \times 1) + (3 \times 2) + (4 \times 3) + (5 \times 2) + (6 \times 1) + (8 \times 1) + (19 \times 1)}{11}$$

$$= 5.727 = 5.7$$

$$M_d = x_{(n+1)/2}$$

$$= x_{(11+1)/2}$$

$$= x_6$$

Thus the answer is 4

1. a)

The n = 8 values are in increasing order of magnitude…

$$Q_1 = x_{(n+1)/4}$$
$$= x_{(8+1)/4}$$
$$= x_{2.25}$$

$$= x_2 + \frac{1}{4}(x_3 - x_2)$$
$$= 7 + \frac{1}{4}(9 - 7)$$
$$= 7.5$$

$$Q_3 = x_{3(n+1)/4}$$
$$= x_{3(2.25)}$$
$$= x_{6.75}$$

$$= x_6 + \frac{3}{4}(x_7 - x_6)$$
$$= 18 + \frac{3}{4}(18 - 18)$$
$$= 18$$

2.

Arrange the n = 10 values in order of size…

1…2…3…4…5…6…7…8…11…12…

$$Q_1 = x_{(n+1)/4}$$
$$= x_{(10+1)/4}$$
$$= x_{2.75}$$

$$= x_2 + \frac{3}{4}(x_3 - x_2)$$
$$= 2 + \frac{3}{4}(3 - 2)$$
$$= 2.75$$

$$Q_3 = x_{3(n+1)/4}$$
$$= x_{3(2.75)}$$
$$= x_{8.25}$$

$$= x_8 + \frac{1}{4}(x_9 - x_8)$$
$$= 8 + \frac{1}{4}(11 - 8)$$
$$= 8.75$$

1.

Arrange the n = 10 values in order as…

8…14…15…16…23…26…28…30…35…61…

Lowest value = 8

Highest value = 61

$$M_d = x_{(n+1)/2}$$
$$= x_{(10+1)/2}$$
$$= x_{5.5}$$

Thus the value is 24.5

$$Q_1 = x_{(n+1)/4}$$
$$= x_{(10+1)/4}$$
$$= x_{2.75}$$

$$= x_2 + \frac{3}{4}(x_3 - x_2)$$
$$= 14 + \frac{3}{4}(15 - 14)$$
$$= 14.75$$

$$Q_3 = x_{3(n+1)/4}$$
$$= x_{3(2.75)}$$
$$= x_{8.25}$$

$$= x_8 + \frac{1}{4}(x_9 - x_8)$$
$$= 30 + \frac{1}{4}(35 - 30)$$
$$= 31.25$$

5 number summary is:

8…14.75…24.5…31.25…61

1.

Arrange the n = 10 values in order

Lowest value = 540
Highest value = 730

$$M_d = x_{(n+1)/2}$$
$$= x_{(10+1)/2}$$
$$= x_{5.5}$$

Thus the value is 565

$$Q_1 = x_{(n+1)/4}$$
$$= x_{(10+1)/4}$$
$$= x_{2.75}$$

$$= x_2 + \frac{3}{4}(x_3 - x_2)$$
$$= 545 + \frac{3}{4}(555 - 545)$$
$$= 552.5$$

$$Q_3 = x_{3(n+1)/4}$$
$$= x_{3(2.75)}$$
$$= x_{8.25}$$

$$= x_8 + \frac{1}{4}(x_9 - x_8)$$
$$= 590 + \frac{1}{4}(650 - 590)$$
$$= 605$$

5 number summary is:
540…552.5…565…605…730