My hypothesis is that the top 3 sets (A, B and C) predict both the size of the angles and the lengths of the line better than the middle 3 sets (D, E and F) and the bottom sets (G, H and I). I think this is correct because the top set is supposed to be smarter and I think they should be able to estimate closer to the correct amount.

The are 185 pieces of data in the whole population.
77 are in the top sets.
70 are in the middle sets.
38 are in the bottom sets.

Overall, there is too much data so I need to choose a sample. I need to make sure my sample size is appropriate. 10 will be too small and is not enough to represent my population and 100 will be too large. I will choose a sample size of 40 as that is not too small and not too large. It is just over 20% of my population so it is enough to get a representative amount of data.

There are not the same amount of people in top, middle and bottom sets so I cannot have the same amount of people from each group. They have to have the same proportion. I worked this out in the table below:

Set	How many people in total	Calculation	Answer	Sample size
Top sets (A, B, C)	77	40*(77/185)	16.64	17
Middle sets (D, E, F)	70	40*(70/185)	15.14	15
Bottom sets (G, H, I)	38	40*(38/185)	8.22	8

Now I know how many people I am going to take from each set, I need a random way of picking people so everyone has a fair chance of being picked. I knew of the random key on the calculator and decided that that was the fairest way of picking the data but first I wanted to see how the random key worked.

When I kept clicking random on the calculator again and again, I noticed two things. Firstly, all the numbers I got were $0 \leq n < 1$ and secondly all the numbers went up to 3 decimal places. This means I can get 0-0.999 so there are 1,000 random numbers (including 0).

As I will get a lot of decimal places I need to choose a way of making them whole numbers. There are two ways I could do this. These are

- I could truncate
- I could round the numbers.

Truncating would mean I would take the whole number and cut of the decimal places. I realised that this will not be very useful as I will never get the highest number.
E.g. If there are 12 numbers, the highest random number is 0.999 and if I multiply that by 12 I get 11.988 and that will count as 11 if I am truncating. This makes it impossible to get 12 so it will not be fair as all the data will not have a fair chance of being picked.

That leaves me with the other method, rounding. Does that give every number the same chance of being picked? I tested this for the top sets sample. I have to choose out of 77 people for my random size of 17. There are 1000 random numbers so I want

to see if all 77 numbers have the same chance of being picked. 1000/77= 12.987…. This calculation shows that not all 77 numbers have the same chance of being picked. That means that if the number you are processing is not a factor of 1000 it is not entirely random. This is shown on the table below:

.

Number we are trying to get	Values of random number that allows us to get that number	How many random numbers	Number gotten with the random number that allows us to get the number we want by rounding
77	0.994-0.999	5	76.538-76.923
76	0.981-0.993	12	75.537-76.461
75	0.968-0.980	12	74.536-75.46
74	0.955-0.967	12	73.535-74.459
70	0.903-0.915	12	69.531-70.455
60	0.772-0.785	13	59.444-60.445
50	0.642-0.655	13	49.434-50.435
40	0.512-0.525	13	39.424-40.425
10	0.124-0.136	12	9.548-10.472
1	0.007-0.019	12	0.539-1.463
0	0-0.006	6	0-0.462

The numbers from 1 to 76 are 12 or 13 random numbers assigned to them so it is relatively fair. However 77 has only 5 numbers assigned to it as 0 has 6. We do not need 0 as one of our results so if the number rounds to 0 we can just say it is 77 so 77 has 11 random numbers assigned to it and makes it fairer.

Top set sample

These are my chosen sample of 17 people from the top set.

Set	D.O.B	Form	Sex	KS3	KS3 Points	Predicted angle	Predicted length
A	07.04.87	10/MA7	F	7	39	60	70
A	31.03.87	10/MA7	F	7	41	50	84
A	11.05.87	10/MA3	M	7	41	72	75
A	11.10.86	10/IT3	F	7	39	60	75
A	14.12.86	10/IT1	F	7	41	60	65
B	17.06.87	10/GG1	F	6	41	50	65
B	30.07.87	10/MA9	F	6	37	60	63
B	16.01.87	10/SC6	M	6	39	87	56
B	25.12.86	10/MA7	F	6	41	67	70

B	27.02.87	10/IT3	F	7	41	70	50
B	05.06.87	10/MA5	M	6	39	80	50
B	30.04.87	10/MA5	M	6	41	74	55
C	31.08.87	10/HI1	F	6	37	60	60
C	12.12.86	10/GG1	F	6	37	55	60
C	29.08.87	10/IT1	F	6	37	65	52
C	26.02.87	10/IT1	F	6	37	55	80
C	28.11.86	10/IT3	M	6	37	55	65

Middle Sets Sample

1000/70=14.29 which is not a whole number so it is not 100% fair. These are my chosen sample of 15 people for the middle set.

Set	**D.O.B**	**Form**	**Sex**	**KS3**	**KS3 Points**	**Predicted angle**	**Predicted length**
D	02.03.87	10/MA5	F	6	39	63	62
D	29.06.87	10/IT1	M	6	39	60	61
D	02.12.86	10/IT3	M	6	37	75	60
D	04.07.87	10/SC6	F	6	35	55	65
D	11.07.87	10/GG1	M	5	37	60	50
D	05.04.87	10/SC6	M	6	35	66	64
D	01.09.86	10/IT3	M	6	41	45	75
E	01.02.87	10/MA5	M	5	37	70	30
E	23.08.87	10/MA7	M	5	31	65	55
E	05.02.87	10/GG1	F	5	31	58	61
E	18.07.87	10/IT1	F	6	37	55	40
E	10.09.86	10/IT1	F	5	33	60	70
F	07.06.87	10/MA7	F	5	33	53	57
F	06.02.87	10/IT1	M	5	27	50	45
F	15.07.87	10/HI1	F	5	33	100	70

Bottom sets Sample

1000/38=26.32 which is not a whole number so it is not 100% fair. These are my chosen 8 people for the bottom set.

Set	**D.O.B**	**Form**	**Sex**	**KS3**	**KS3 points**	**Predicted angle**	**Predicted length**
G	07.07.87	10/IT1	M	4	33	30	60
G	03.05.87	10/HI1	F	4	29	60	40
G	21.05.87	10/GG1	M	5	37	60	60
G	25.11.86	10/MA5	M	4	31	60	52
H	12.11.86	10/MA3	F	4	31	75	75
H	24.04.87	10/SC6	F	4	21	65	65
I	21.05.87	10/MA9	F	3	23	50	60
I	25.07.86	10/MA7	M	3	23	55	90

Now I have chosen my sample data, I need to start comparing them. I have put the information on stem and leaf diagrams so it will be easier to read and work out the averages.

Stem and Leaf for angles

From the stem and leaf diagrams I can calculate the mean, mode and range. The mean is the arithmetic average; the sum of the data divided by the sample size. One problem with using the mean is that it does not often show the typical outcome. If there is one outcome that is very far from the rest of the data, then the mean will be affected by this outcome.

The median is a measure of the central tendency of a data set. It is the middle value in a data set, when the values are ranked from lowest to highest. The median is better for describing the typical value.

The mode is the single class in a statistical distribution having the greatest frequency. The mode shows what most people guessed.

The range shows the difference between the minimum value and the maximum value in a set of data. The range helps identify best and worst case and process variability.

Top set

Mean = sum of all numbers = 64
17

Mode = 60

Median = 60

Range = 87-50 = 37

Top set

3	
4	
5	00555
6	0000057
7	024
8	07
9	
10	

Middle set

Mean = sum of all numbers = 72
15

Mode = 70

Median = 70

Range = 100-55 = 45

Middle set

3	
4	
5	5
6	03558
7	000356
8	05
9	
10	0

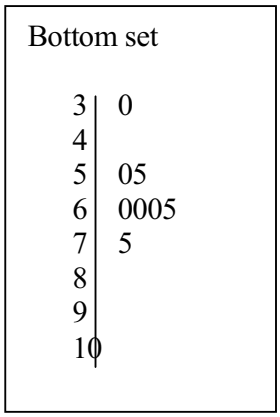
Bottom set

Mean = sum of all numbers = 57
8

Mode = 60

Median = 60

Range = 75-30 = 45



To make my results clearer to read I have condensed them into the table below:

	Mean	**Mode**	**Median**	**Range**
Top sets	64	60	60	37
Middle sets	72	70	70	45
Bottom sets	57	60	60	45
Actual value	64			
Overall mean value	64.7			

This table can help us to compare the data. The top set mean is less than the overall mean and is the actual value. It is closer to the actual value than the other means so I am on the right track with my hypothesis. The middle set mean too large and is 8 away from the actual value. The bottom set is too small and is 7 away from the actual value so the bottom set had a better mean than the top set.

Top and bottom set have a mode of 60 which is 4 away from the actual value and better than the middle set mode of 70 which is 6 away from the actual value. This means that more people thought the size of the angle was 60 and 70.

As I said before, the median shows a typical value. Surprisingly the top and bottom set also have the same median on 60 again and the middle set is further off with a median of 70. So far it seems as though the bottom set predicted angles better than the middle set.

The range shows how spread out the data is. The top set has the smallest range of 37 than the middle and bottom sets which have a range of 45. This means that the top sets results had less spread than the other two sets.

The top set had the best mean and range. It was joint with the bottom group with the best mode and median so it is fair to say that top set were better at estimating angles than the bottom and middle sets so far.

The surprising this that it seems as thought the bottom set are better than the middle set by the means I have been using to compare so far. Now I will look at the lengths

Stem and leaf for lengths

I will do stem and leaf for the lengths now. From the stem and leaf diagrams I can calculate the mean, mode and range.

Top set

Mean = sum of all numbers = 64
17

Mode = 65

Median = 65

Range = 84-50 = 34

Top set

3	
4	
5	00256
6	003555
7	0055
8	04
9	
10	

Middle set

Mean = sum of all numbers = 58
15

Mode = 61, 70

Median = 61

Range = 75-30 = 45

Middle set

3	0
4	05
5	057
6	011245
7	005
8	
9	
10	

Bottom set

Mean = sum of all numbers = 63
8

Mode = 60

Median = 60

Range = 90-40 = 50

Bottom set

3	
4	0
5	2
6	0005
7	5
8	
9	0
10	

To make my results easier to read I have condensed them into the table below:

	Mean	**Mode**	**Median**	**Range**
Top sets	64	65	65	34
Middle sets	58	61, 70	61	45
Bottom sets	63	60	60	50

Actual value	59
Overall mean value	61.1

For the lengths top set results were not as good as they were for the angles. The mean for top set is the most furthest away from the actual value and the middle set is the closest. Even the bottom set got a closer mean than top set.

All the modes were larger than the actual value but bottom set was only 1 away. The middle set had two modes so we cannot really get any information from that as the two modes are far apart. Top sets mode was off the actual value by 6.

The medians were also all larger than the actual value and again bottom set has the closest median to the actual value and was only off by 1. Top sets median was the most far- off of the three sets as middle set was closer than top. On the other hand, top set got the smallest range so the data was less spread apart.

Box Plots

To do a box plat I need to use some of the information I got from before like the median but I also need to work out the lower and upper quartiles. The formula for them is written below:

Lower quartile	=	$^1/_4\,(n+1)$
Median	=	$^1/_2\,(n+1)$
Upper quartile	=	$^3/_4\,(n+1)$

Top set	
Highest Number	87
Lowest Number	50
Median	60
Lower Quartile	55
Upper Quartile	71

Middle set	
Highest Number	100
Lowest Number	55
Median	70
Lower Quartile	65
Upper Quartile	76

Bottom set	
Highest Number	75
Lowest Number	30
Median	60
Lower Quartile	57
Upper Quartile	62

The information above is enough for me to draw a box plot for the angles.

The bottom set has the smallest inter quartile rage and top set has the largest. However, the actual value is not in bottom set or middle set inter quartile range but it is in top sets. So although top sets inter quartile range is larger, it is more accurate because it is around the actual value. The lower quartile of the middle set is larger than the actual value and the upper quartile of the bottom set is smaller than the actual value. In conclusion, from the box plots, top set estimated the angles better than middle and bottom set. Now I need to do box plots for the lengths.

Top set	
Highest Number	84
Lowest Number	50
Median	65
Lower Quartile	55.5
Upper Quartile	72.5

Middle set	
Highest Number	75
Lowest Number	30
Median	61
Lower Quartile	50
Upper Quartile	65

Bottom set	
Highest Number	90
Lowest Number	40
Median	60
Lower Quartile	54
Upper Quartile	77.5

The information above is what I need to draw my box plots.

The box plots seem better in the lengths than they did in the angles as the actual value is in all three inter quartile ranges. Top set has the smallest range and most of the predictions made were larger than the actual value. The middle set has the largest range so its data is more spread out. Top sets median is the largest and most further away from the actual value. The actual value is closer to the lower quartile of top set and bottom set but is more or less in the middle of middle sets. From these box plots it is difficult to say which set done best.

The mean, mode and median do a nice job in telling where the average of the data is, but often we are interested in more. We need a measure of how far the data is spread apart. This is what standard deviation does.

Standard deviation for angles

Standard deviation is a statistical measure of spread or variability. It is a statistic that measures the dispersion of a sample. This is the formula

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

Σ **- sigma**

$\bar{x}$ **- mean**

Top set

Angle	Angle – Actual value (64) **x**
80	16
67	3
74	10
60	-4
60	-4
50	-14
60	-4
72	8
60	-4
70	6
60	-4
55	-9
55	-9
87	23
55	-9
65	1
50	-14

Middle set

Angle	Angle – Actual value (64) **x**
60	-4
53	-11
65	1
63	-1
60	-4
100	36
70	6
58	-6
55	-9
75	11
55	-9
50	-14
60	-4
66	2
45	-19

Bottom set

Angle	Angle – Actual value (64) **x**
30	-34
60	-4
60	-4
60	-4
75	11
65	1
50	-14
55	-9

	Mean of x	Standard deviation of x
Top set	-0.47	10.02
Middle set	-1.67	12.49
Bottom set	-7.13	12.23

The mean of x showed some interesting results. First of all, top set got a mean of -0.47 which is very small and means that most of the far off negative ones balanced out the far off positive ones very well. It had the mean which was closest to zero so it was the best one. Middle set mean was not too bad but it was further away from zero than top set. Bottom set, on the other hand, got a mean of -7.13 which is not very

good. All the sets got a negative mean which shows they guessed less than the actual value was but bottom set guessed way too low.

The standard deviation of top set is less than the other sets which shows that it had less spread from the actual values than middle set and bottom set. The is not a lot of difference between middle and bottom sets standard deviation although bottom sets is slightly smaller. For angles in terms of standard deviation, top set estimated the best.

Standard deviation for lengths

Length	Length – Actual value (59) **x**
70	11
84	25
75	16
75	16
65	6
65	6
63	4
56	-3
70	11
50	-9
50	-9
55	-4
60	1
60	1
52	-7
80	21
65	6

Length	Length – Actual value (59) **x**
62	3
61	2
60	1
65	6
50	-9
64	5
75	16
30	-29
55	-4
61	2
40	-19
70	11
57	-2
45	-14
70	11

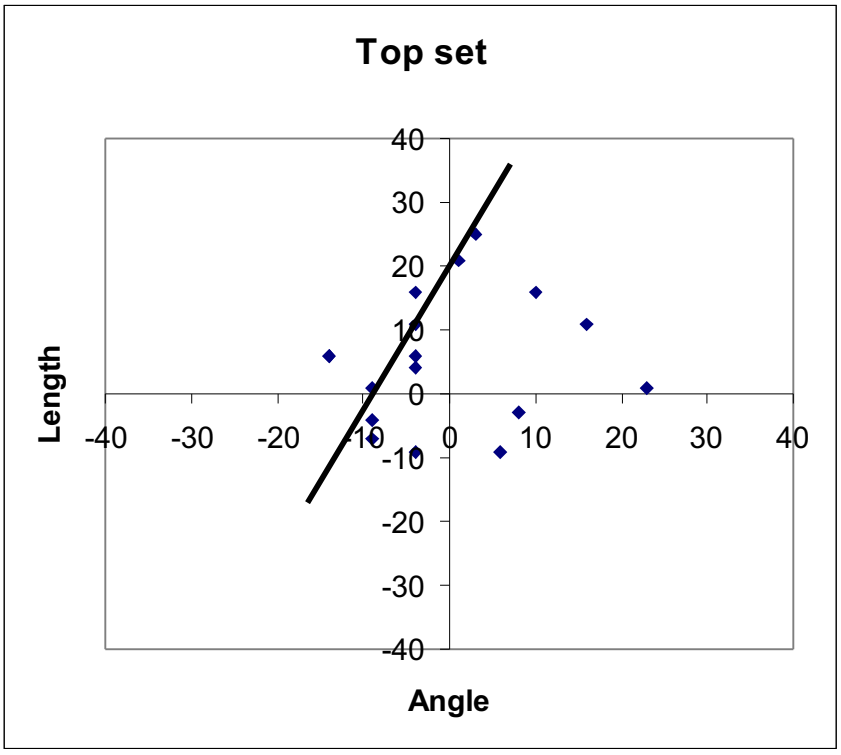
Length	Length – Actual value (59) **x**
60	1
40	-19
60	1
52	-7
75	16
65	6
60	1
90	31

	Mean for x	Standard deviation for x
Top set	5.41	9.95
Middle set	-1.33	11.69
Bottom set	3.75	13.94

I wanted the mean to be close to zero and the middle set got the best mean and it was negative. That means overall they guessed the length too small. Top set and bottom set got a positive mean so they guessed the lengths too big. However the top set got the mean most far-off as even the bottom groups mean was smaller than it.

In the standard deviation however, the top set got the smallest so their data was less spread out. The bottom set has the highest standard deviation so its data was the most spread out and few people got the accurate length.. As the standard deviation is more accurate than the mean, top set still got good estimates.

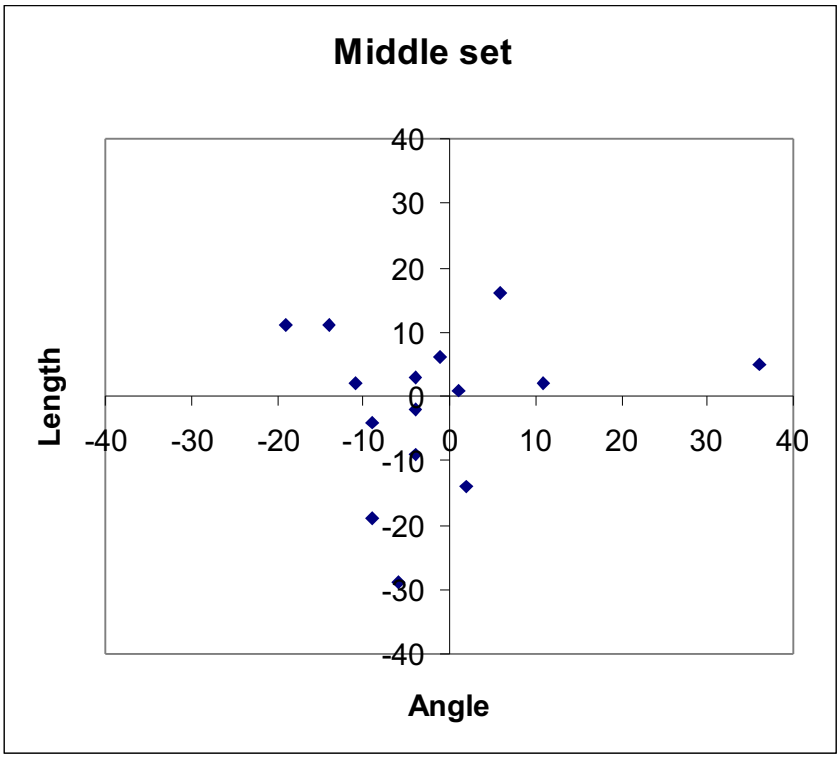
Scatter diagram

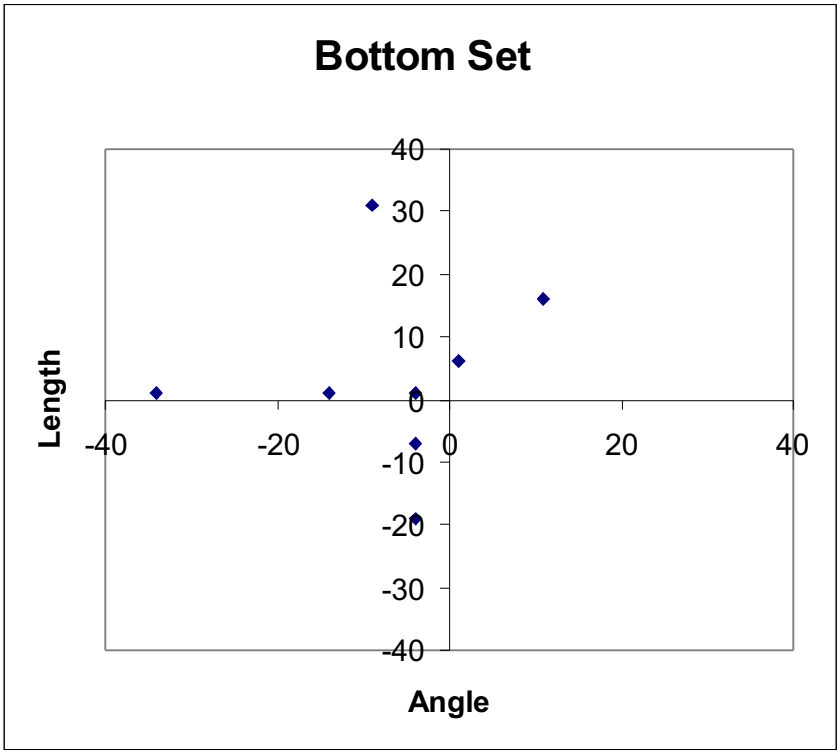


I wanted to see if there was any correlation between peoples estimates on the angles and the lengths. To make this fair I took the **x** value of what I done in the standard deviation which showed either "Angle – Actual value (59)" or "Length – Actual Value (64)". I plotted the angle on the x axis and the length on the y axis.

For the top set, there seems to be a very weak positive correlation between the lengths and angles. I have drawn a line of best fit but 7 out of the 17 pieces of data are very far away from the line of best fit.

I drew another scatter diagram for the middle sets and bottom sets on the following page.





There seems to be no correlation for these scatter diagrams as well. This shows that they estimates the pupils made about the length of the line and the size of the angle were not related.

Conclusion

For angles, my hypothesis was correct as top set had better mean, median, mode and range. It also had a better box plot and standard deviation.

For the lengths of the line however, it was not as simple to see which set predicted the lengths better. Top set had the worst mean but the best standard deviation. Top set didn't get a good mode or median but it had the smallest range. From the box plots, it was impossible to see which set done better.

So overall my hypothesis was right for the size of the angle but not for the length of the line.