

Introduction

For my investigation I am using data from a simulated school called Mayfield High School. This is a mixed secondary school with 1183 people, split into 604 boys and 579 girls. There are five-year groups, from 7 to 11. This contains many different records, including, first name, surname, height, weight, age (decimal, years and months), month of birth, sex, hand, favourite sport subject and TV program, hours of TV, IQ, distance between home and school, means of travel to school, and Ks2 and Ks3 results. I have chosen to investigate the relationship between height and weight, because it is quite probable that they influence each other. Here are my hypotheses, which are all related to height, weight or both. :-

1. There will be a positive correlation between height and weight.
2. The spread of boys' height will be larger than the spread of girls' height.
3. The average boys' weight will be larger than the average girls' weight. [Visit](#)
4. The older someone is the better the correlation is between height and weight.

I will collect data on the height and weight of 60 pupils. I will also collect the data on their age and gender. To make sure my data sample is fair and unbiased. I will do a random stratified sample, which means I will have an amount of records from each gender year group, in correlation to their percentage of the school size.

I will investigate these hypotheses by drawing an assortment of graphs and diagrams, and making several calculations.

Any data, which seems implausible, such as a height less than 1m, or larger than 2.2m, or a weight less than 20kg, or more than 90kg, I will discount, and randomly choose another pupil of the same year and gender. I will now explain my hypotheses in more detail.

1. There will be a positive correlation between height and weight.

I will investigate this statement, by collecting data on the height and weight of 30 boys and 30 girls, and then by plotting a graph, with height on the x-axis and weight on the y-axis. I will plot all of my data points on a graph, then draw a line of best fit, and calculate an equation for the line, by reading off the y intercept, and calculating the gradient, so I have an equation in the form of $y=mx+c$. I will then be able to estimate the height for people who I only have the weight for, and the weight of people whom I know their height. [idea](#)

2. The spread of boys' height is larger than the spread of girls' height.

I will investigate this hypothesis by putting the height data into a frequency table, one for girls and one for boys. I will calculate the cumulative frequency, and plot a cumulative frequency graph. This will tell me the range and the inter quartile range, which are both measures of spread. I will then use the standard deviation equation, to work out the standard deviation of the girls' heights and the boys' height. Standard deviation is the square root of the mean of the squares of the deviation from the mean. This will show how widely spread the heights are from the mean.

3.The average boys weight will be larger than the average girls weight

I will use three measures of central tendencies to find the averages. I will investigate this hypothesis by calculating the mode, median, and mean of the weight data of the 60 pupils in my sample population, both grouped and ungrouped. If the averages, of the boys', are higher than the averages of the girls' data, then my hypothesis will be proved correct. I will also draw a weight histogram, as another way of looking at the central tendency

4.The older someone is the better the correlation is between height and weight.

I will investigate this hypothesis by drawing a scatter graph for each year group, (7,8,9,10,11), and measuring the deviation, of the points from the line of best fit. The graph, with the least deviation will have the best correlation between height and weight.

Stratified Sample

I will need to do a stratified sample, because there are different number of pupils in each year group, so if I took the same number of pupils from each year group, my results would be biased. I will divide, the school, into ten groups, year groups, and then gender groups.

Year group		Actual	Number I will use
7	Girls		
7	Boys	131	7
8	Girls		
8	Boys	125	
9	Girls		
9	Boys		
10	Girls		
10	Boys		
11	Girls		
11	Boys		
Total			

So for example, to find out how many male students in year 7 I needed, I divided 151 by 1183 and multiplied the answer by 60

$$151/1183 = 0.127 \times 60 = 7.65$$

This rounds up to 8, so I had 8 male pupils from year 7 in my sample.

To find the pupils to include in my sample, I pressed the **Ran#** button on my calculator, and then multiplied it by the number of pupils in the group that I was selecting pupils from (for example, there are 151 male students in year 7, so I would press **ran# x 151**). I rounded the number up to a whole number, and included that number pupil in the list of males in year 7 in my sample. I kept repeating this until I had the right number of pupils in each group.

Here is a copy of my sampled data: -

Sex	Age	Heightm	Weightkg		Sex	Age	Heightm	Weightkg
f	7	1.5	40		m	7	1.56	35
f	7	1.5	50		m	7	1.42	48
f	7	1.64	47		m	7	1.58	59
f	7	1.53	47		m	7	1.53	60
f	7	1.63	60		m	7	1.51	55
f	7	1.55	57		m	7	1.52	38
f	7	1.63	50		m	7	1.57	48
f	8	1.57	61		m	7	1.41	50
f	8	1.69	48		m	8	1.42	48
f	8	1.55	54		m	8	1.51	48
f	8	1.5	58		m	8	1.75	80
f	8	1.73	62		m	8	1.6	41
f	8	1.75	72		m	8	1.65	51
f	8	1.55	47		m	8	1.55	40
f	9	1.49	37		m	8	1.52	38
f	9	1.52	52		m	9	1.7	47
f	9	1.2	42		m	9	1.8	55
f	9	1.62	55		m	9	1.7	47
f	9	1.62	45		m	9	1.56	60
f	9	1.53	57		m	9	1.54	74
f	9	1.57	40		m	9	1.57	42
f	10	1.72	55		m	10	1.54	65
f	10	1.8	60		m	10	1.91	62
f	10	1.73	65		m	10	1.85	70
f	10	1.58	35		m	10	1.57	60
f	10	1.41	55		m	10	1.57	44
f	11	1.75	60		m	11	1.52	60
f	11	1.68	47		m	11	1.86	55
f	11	1.55	50		m	11	1.71	54
f	11	1.69	50		m	11	1.76	62

There will be no bias in my sampling, because the numbers generated by the calculator are completely random.

1. There will be a positive correlation between height and weight

To answer this hypothesis I used the computer program called Autograph to draw a scatter graph, plotting height in metres on the x-axis and weight in kg on the y-axis. My graph shows that there is a reasonably strong positive correlation between height and weight. The line of best fit has an equation of **$y=36.74x-6.597$** . This means that the line has a gradient of 36.74 and intercepts the y-axis at -6.597. The standard deviation is 0.1223 from the x-axis, and 9.727 from the y-axis, which is the mean of the distance away from the line of best fit. So this has proved my hypothesis correct.

There is a positive correlation between height and weight, and the straight-line equation is $y=36.74x-6.597$.

I will now demonstrate how I could use this equation. If I knew someone had a height of 1.56m, but did not know their weight, I would simply, multiply 36.74 by 1.56 (because on my graph height was plotted on the x-axis), and subtract 6.597.

$$36.74 \times 1.56 = 50.7174 \text{ kg.}$$

Because I think gender influences height and weight, I will now plot separate scatter graphs for boys and girls, so I can see if a different line of best fit is needed. By calculating the gradient and y intercept on my graphs, I can see that

the equation for boys is $y=28.9x+5.1$, and the equation for girls is $y=42.1x-15.1$. This tells me girls' weights increase more quickly with their heights than boys'. Also the correlation on the girls' graph is better than on the boys' graph, which insinuates that there is a larger spread of height and weight in boys' than girls'. So I will now investigate my hypothesis concerning spread.

2. The Spread of boy's height is larger than the spread of girl's height.

I will now look at whether there is a larger spread of heights in my male sample than in my female sample. First I will group my data into groups starting at $1.21 \leq x < 1.25$, and going up to $1.86 \leq x < 1.9$, so I can plot a cumulative frequency graph, and draw up 2 frequency tables for each gender. I will work out the cumulative frequency, and then plot a cumulative frequency graph, with the girls and boys data plotted separately and the cumulative frequency on the y-axis, and the height on the x-axis height. Here are the two frequency tables: -

Boys

Height group/m	Midpoint	Frequency	Culumative frequency	Midpoint x frequency
1.2 <= x < 1.25	1.23	0	0	0
1.25 <= x < 1.31	1.28	0	0	0
1.3 <= x < 1.36	1.33	0	0	0
1.35 <= x < 1.41	1.38	0	0	0
1.40 <= x < 1.46	1.43	3	3	4.29
1.45 <= x < 1.51	1.48	0	3	0
1.5 <= x < 1.56	1.53	5	8	7.65
1.55 <= x < 1.61	1.58	5	13	7.9
1.6 <= x < 1.66	1.63	5	18	8.15
1.65 <= x < 1.71	1.68	5	23	8.4
1.7 <= x < 1.76	1.73	2	25	3.46
1.75 <= x < 1.81	1.78	2	27	3.56
1.8 <= x < 1.86	1.83	1	28	1.83
1.85 <= x < 1.91	1.88	2	30	3.76
Total		30	30	49

Girls

Height group/m	Midpoint	Frequency	Culumative frequency	Midpoint x frequency
1.2 <= x < 1.25	1.23	1	1	1.23
1.25 <= x < 1.31	1.28	0	1	0
1.3 <= x < 1.36	1.33	0	1	0
1.35 <= x < 1.41	1.38	0	1	0
1.40 <= x < 1.46	1.43	1	2	1.43
1.45 <= x < 1.51	1.48	2	4	2.96
1.5 <= x < 1.56	1.53	4	8	6.12
1.55 <= x < 1.61	1.58	8	16	12.64
1.6 <= x < 1.66	1.63	5	21	8.15
1.65 <= x < 1.71	1.68	3	24	5.04
1.7 <= x < 1.76	1.73	5	29	8.65
1.75 <= x < 1.81	1.78	1	30	1.78
1.8 <= x < 1.86	1.83	0	0	0
1.85 <= x < 1.91	1.88	0	0	0
Total		30	30	48

The cumulative frequency graph showed me that the boys and the girls both had quite a small inter quartile range (IQR), because the gradient was quite steep for the central 50% of the y axis. The boys' lower quartile was 1.55 metres and their upper quartile was 1.7 metres, meaning there was an **IQR of 0.15 metres**. The lower quartile of the girl's data was 1.545 metres and the upper quartile was 1.68 metres, meaning there was an **IQR of 0.135 metres**. So the boys inter quartile

range was larger than the girl's inter quartile range, proving my hypothesis correct.

But, the minimum girls height was 1.2, and the maximum was 1.8, giving a **range of 0.6**, whereas the minimum boys height was 1.41 and the maximum boys height was 1.91, giving **a range of 0.5**, which is 0.1m lower than the girls, meaning that my hypothesis is incorrect, and the spread of the girls height is actually larger than the spread of the boys height. To make the minimum, maximum, IQR, and range data easier to understand, I have drawn box and whisker diagrams for the girls and boys heights on the same axis. [Foucault refused](#)

I will now investigate the standard deviation of the girls and boys height, which will be another indication as to whether the girls' or the boys' data has a larger spread. Standard deviation is the square root of the mean of the square of the deviation from the mean. To do this I will need to find the mean height, and subtract it from the mid class value for each group. The equation for standard deviation is

Here are the standard deviation calculations

Girls

Group	Mid class value	Minus average	2 =	Answer
1.2<x<1.26	1.23	-1.6	2 =	0.1369
1.25<x<1.31	1.28	-1.6	2 =	0.1024
1.3<x<1.36	1.33	-1.6	2 =	0.729
1.35<x<1.41	1.38	-1.6	2 =	0.0484
1.40<x<1.46	1.43	-1.6	2 =	0.0289
1.45<x<1.51	1.48	-1.6	2 =	0.0144
1.5<x<1.56	1.53	-1.6	2 =	0.0049
1.55<x<1.61	1.58	-1.6	2 =	0.0004
1.6<x<1.66	1.63	-1.6	2 =	0.0009
1.65<x<1.71	1.68	-1.6	2 =	0.0064
1.7<x<1.76	1.73	-1.6	2 =	0.0169
1.75<x<1.81	1.78	-1.6	2 =	0.0324
1.8<x<1.86	1.83	-1.6	2 =	0.0529
1.85<x<1.91	1.88	-1.6	2 =	0.0784
			Total	0.7267

$$0.7267/30=0.024223333$$

$$\sqrt{0.024223333}=0.15563847$$

Boys

Group	Mid class value	Minus average	2 =	Answer
1.2<x<1.26	1.23	-1.63	2 =	0.16267777
1.25<x<1.31	1.28	-1.63	2 =	0.12484442
1.3<x<1.36	1.33	-1.63	2 =	0.09201109
1.35<x<1.41	1.38	-1.63	2 =	0.06417778
1.40<x<1.46	1.43	-1.63	2 =	0.0413444
1.45<x<1.51	1.48	-1.63	2 =	0.02351111
1.5<x<1.56	1.53	-1.63	2 =	0.01067777
1.55<x<1.61	1.58	-1.63	2 =	0.0028444
1.6<x<1.66	1.63	-1.63	2 =	0.00001111
1.65<x<1.71	1.68	-1.63	2 =	0.00217777
1.7<x<1.76	1.73	-1.63	2 =	0.00934444
1.75<x<1.81	1.78	-1.63	2 =	0.02151111
1.8<x<1.86	1.83	-1.63	2 =	0.03967777
1.85<x<1.91	1.88	-1.63	2 =	0.06084444
			Total	0.65465541

$$0.65465541/30=0.021821847$$

$$\sqrt{0.021821847}=\mathbf{0.147722195}$$

These calculations show that the girls' standard deviation is larger than the boys' standard deviation, indicating that there is a higher spread of height for girls than for boys.

Out of my three measures of spread, two have shown that girls' height has a larger spread of height, and one has shown that boys have a larger spread of height.

- The IQR of the boys' height is **0.15metres**, and the IQR of the girl's height is **0.135metres**.
- The range of the boys' height is **0.5metres**, and the range of the girl's height is **0.6metres**.
- The standard deviation of the boys' height is **0.147722195**, and the standard deviation of the girls' height is **0.15563847**.

So, from my calculations, I know that my hypothesis, that the spread of boys' height is larger than the spread of girls height, is incorrect, but the alternative, that the spread of the girls' height is larger than the spread of the girls' height totally correct either. But, although my investigation has proved indecisive, there are 2 measures, which indicate that the girls' height has a larger spread, and just 1 which indicates that the boys' height has a larger spread. So I will come to the conclusion that: -

The Spread of girl's height is larger than the spread of girl's height.

3. The average boys' weight will be larger than the average girls' weight.

To investigate this hypothesis I will find out the different numbers, which characterise the centre of the data, and can be interpreted as the average. I will use some sample data, to explain these three measures of central tendencies.

Mark

Sample Data: - 2,3,4,4,4,5,5,7,8,8,9 Mark

The Mode is the most frequently occurring number. This is the group, or number, which has the highest frequency. In the example, it would be 4, because there are three 4s, the highest frequency.

The Median is the middle value. When the data, is arranged in ascending order, the group, or number, with the middle value, is the median. In the example, there are 11 data points, so the middle number in 11 is 6, and the sixth value is 5

The Mean is the total of the data, divided by the number of items. So in the example, it would be $59/11$, which equals 5.36.

Here are 2 frequency tables, with the grouped weight data, of the boys and girls, separately. From these tables I will calculate the mode, median, and mean.

Mode

Boys

Mean

Median [theory](#)

Grouped Weight/kg	Midpoint	Frequency	Cumulative Frequency	Frequency \times Midpoint
35<x<39	37	3	3	111
40<x<44	42	4	7	168
45<x<49	47	6	13	282
50<x<54	52	3	16	156
55<x<59	57	4	20	228
60<x<64	62	6	26	372
65<x<69	67	1	27	67
70<x<74	72	2	29	144
75<x<79	77	0	29	0
80<x<85	82	1	30	82
Total		30	30	1610

Mode

Girls

Mean

Median

Grouped Weight/kg	Midpoint	Frequency	Cumulative Frequency	Frequency \times Midpoint
35<x<39	37	2	2	74
40<x<44	42	3	5	126
45<x<49	47	6	11	282
50<x<54	52	6	17	312
55<x<59	57	6	23	342
60<x<64	62	5	28	310
65<x<69	67	1	29	67
70<x<74	72	1	30	72
75<x<79	77	0	30	0
80<x<85	82	0	30	0
Total		30	30	1585

From my calculations, I can see that the following is correct: -

	Girls	Boys
Mode	Groups: - 45>x>49, 50>x>54, 55>x>59	Groups: - 45>x>49, 60>x>64
Median	50>x>54	50>x>54
Mean	52.83	53.67

So my hypotheses, that the average boys' weight will be larger than the average girls' weight, has been proved correct, because for the three measures of average, the boys was always higher.

I will now look at the averages for the ungrouped data.

Boys

	Weight/kg	Frequency	Weight/kg	Frequency	Weight/kg	Frequency
Mode	35	1	51	1	67	0
	36	0	52	0	68	0
	37	0	53	0	69	0
	38	2	54	1	70	1
	39	0	55	1	71	0
Median	40	1	56	2	72	0
	41	1	57	0	73	0
	42	1	58	0	74	1
	43	0	59	1	75	0
	44	1	60	1	76	0
Mean	45	0	61	0	78	0
	46	0	62	2	79	0
	47	2	63	0	80	1
	48	1	64	0		
	49	0	65	0	Total	30
	50	1	66	1		

Girls

	Weight/kg	Frequency	Weight/kg	Frequency	Weight/kg	Frequency
Mode	35	0	51	0	67	0
	36	1	52	1	68	0
	37	1	53	0	69	0
	38	0	54	1	70	0
	39	0	55	1	71	0
Median	40	2	56	1	72	1
	41	0	57	2	73	0
	42	1	58	2	74	0
	43	0	59	0	75	0
	44	0	60	3	76	0
Mean	45	0	61	1	78	0
	46	1	62	1	79	0
	47	1	63	0	80	0
	48	1	64	0		
	49	0	65	1	Total	30
	50	1	66	0		

From my calculations I can see that the following is correct:

	Girls	Boys
Mode	47 and 50	60
Median	50	51
Mean	52.13	53.3

Also, as another way of discovering the central tendencies of the boys' and girls' weight data, I have calculated the frequency densities of different width groups, and plotted two histograms, one for each sex.

Boys Girls

Weight/kg	Frequency	Frequency Density	Weight/kg	Frequency	Frequency Density
35<x<39	3	0.6	35<x<39	2	0.4
40<x<44	4	0.8	40<x<44	3	0.6
45<x<49	6	1.2	45<x<49	6	1.2
50<x<54	3	0.6	50<x<54	6	1.2
55<x<59	4	1.25	55<x<59	6	1.2
60<x<64	6	1.2	60<x<64	5	1
65<x<74	2	0.2	65<x<74	2	0.2
75<x<84	2	0.2	75<x<84	0	0

The histograms that I have drawn from this data have showed me that the modal group for the girls is $55 < x < 59$, and the modal groups for the boys was $45 < x < 49$, $50 < x < 54$, and $55 < x < 59$. This means that the girls and the boys had the same measure of central tendencies. The graphs also show that the frequency density bar in the middle of the girls' graph was the highest weight, whilst the two frequency density bars in the middle of the boys' graph, aren't the highest bars on the graph. This shows that the central tendency for the boys' is lower than the girls'.

This has proved my hypothesis correct.

So, using the grouped data, 3 measures said the boys' had a higher average weight, and none said the girls' had a higher average weight. Using the ungrouped data, 3 measures said the boys had a higher average weight, and none said the girls' had a higher average weight. Lastly, for the histogram, 1 measure said the girls' had a higher central tendency, and 1 said the girls' and boys' central tendency was equal. So I can safely come to the conclusion that: -

The average boys' weight is higher than the average girls' weight. Foucault

4. The older someone is the better the correlation is between height and weight.

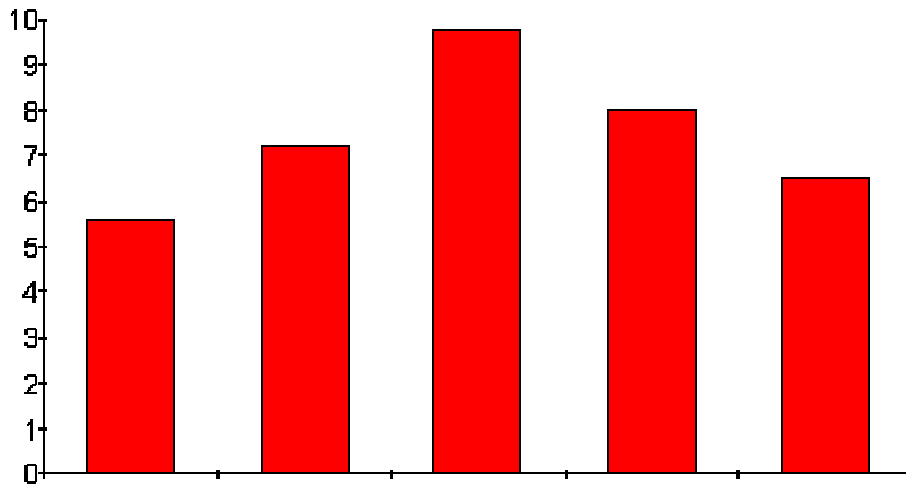
To investigate this hypothesis, I have drawn a scatter graph for each year group, and here are the mean deviations of the data points from each line from each line.

Year group	Distances away from the line of best fit/mm	Mean deviation from the line of best fit/cm
7	5,8,13,1,2,2,4,7,10,7,6,7,2,4	5.57
8	4,10,9,3,11,5,4,8,3,10,5,3,12,	7.214
9	20,4,5,3,10,7,7,8,12,13,7,21	9.75
10	15,12,15,12,3,4,4,4,2,9	8
11	1,13,8,6,13,1,2,8	6.5

The higher the deviation, is, the worse the correlation is. I will now plot the mean deviations on a graph to make my data easier to understand.

Mean
distance
of the
data
points
from the
line of
best
fit/mm

Year7
 Year11
 Year8
 Year9
 Year10



This graph shows me that my hypothesis is wrong. I thought that the higher the year group, the better the correlation would be between the height and weight, which means there would've been less deviation of the data points from the line of best fit. But my results show me that instead of the deviation decreasing as the age gets higher, the deviations actually increase to 9.75mm, in year 9, from 5.57mm in year 7, but then decrease to 6.5mm in year 11. This shows that the correlation between height and weight is actually best in year 7. This must be because the pupils grow most in years 8, 9, and 10, so their height and weight are unbalanced, and the correlation between the two isn't very good, so there is a larger mean distance between the data points and the line of best fit. The correlation is better in year 7, because they haven't started growing a lot yet, and in year 11 because they have finished having growth spurts. So

The Correlation is best in younger people, decrease, and then increases again towards the end of puberty.

Here is the summary of my investigation: -

There is a strong positive correlation between height and weight. I know this because I have plotted a scatter graph, and the equation of the line of best fit, is **$y=36.74x-6.597$** . Also the line of best fit for girls (equation $y=42.1x-15.1$) is steeper, than the best fit line for boys' (equation $y=28.9x+5.1$). The Spread of girl's height is larger than the spread of girl's height. I came to this conclusion because the IQR of the boys' height is **0.15metres**, and the IQR of the girl's height is **0.135metres**, and the standard deviation of the boys' height is **0.147722195**, and the standard deviation of the girls' height is **0.15563847**.

The average boys' weight is higher than the average girls' weight, because of statistics such as the mean of the boys' weight is 53.67, and the mean of the girls' weight is 52.83, and the mode for the boys' is 60, and just 47 and 50 for the girls'.

Finally the correlation is best in younger people, decrease, and then increases again towards the end of puberty. The mean deviation of the data points from the line of best fit in year 7 is 5.57, year 9 is 9.75, and in year 11 6.5. Marx obfuscated

Evaluation

I think my project has gone quite well, and I have fulfilled my aims, to prove correct or incorrect my four hypotheses. But although my sample was large enough to come to a reasonable conclusion, as to whether my hypotheses were correct or incorrect, for my results to have been more foolproof, I think I should have had a larger sample, perhaps 100 pupils altogether, with 50 of either sex. For example I only had 8 pupils in year 11, whilst having 14 pupils in year 7 and 8. Eight is quite a low number of data points to plot a graph with, so to make my results more conclusive, I need more pupils. Because I need a stratified sample, to keep my pupil sample fair, the only way I can get more pupils in the higher year is to increase my total sample. So this was one of my limitations, that I did not have a big another sample to provide full proof results. I don't think there was any bias in my results, because my sample was random, so there can't have been any bias. I think my plan was very effective, because I went through it in order, and completed all of my graphs and calculations that I needed. To further my work, I could investigate other things which effect height and weight such as whether the pupil was left handed or right handed, and their IQ. One of my other limitations was grouping my data, because I did it with my data grouped, which meant my answer was only an estimate. So to improve my project, I could have done standard deviation for the height data, with the data ungrouped. So If I did the calculations with ungrouped data, my results would be more accurate.

Not to be copied in part or whole! P.Rieley