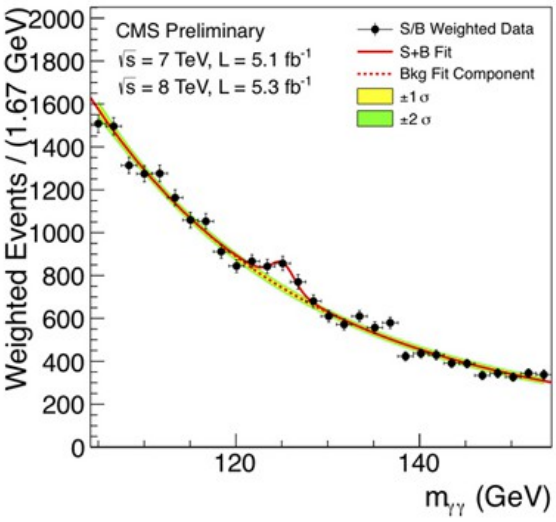
Particle Discovery – (Data Analysis)

A major component of being a physicist (or any scientist) is being able to collect data and then analyse it. In this exercise, you will analyse data that was collected from the Large Hadron Collider at CERN, a collaboration of thousands of scientists from hundreds of countries. This data shows the energy produced from proton collisions. There is a "bump" in the data this could indicate that a new particle has been discovered. Data similar to this was used to discover the Higgs Boson in 2012 (See diagram to the right). These instructions take you through a basic version of the process scientists use to prove the existence of new particles.



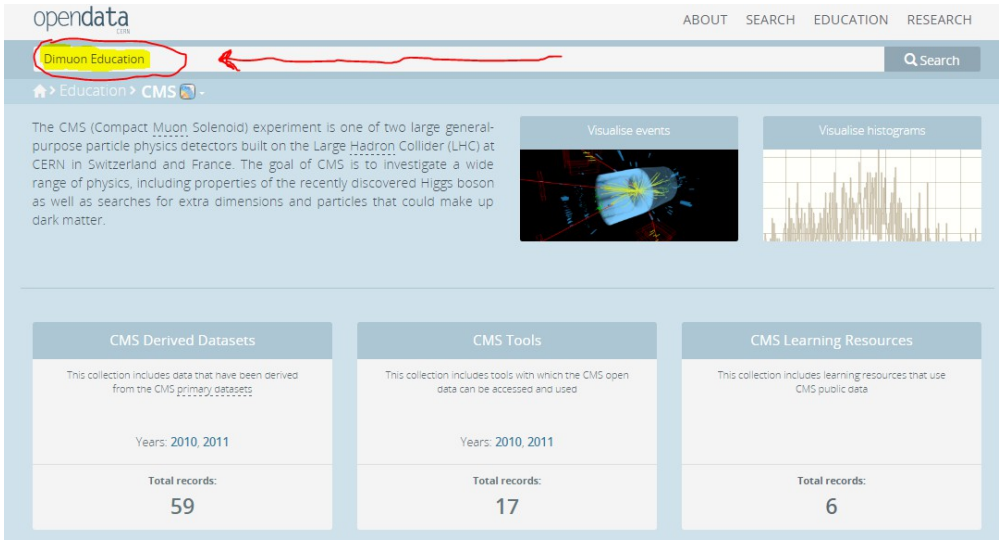
1.) Download and Open "Dimuon.csv".

Go to: http://opendata.cern.ch/

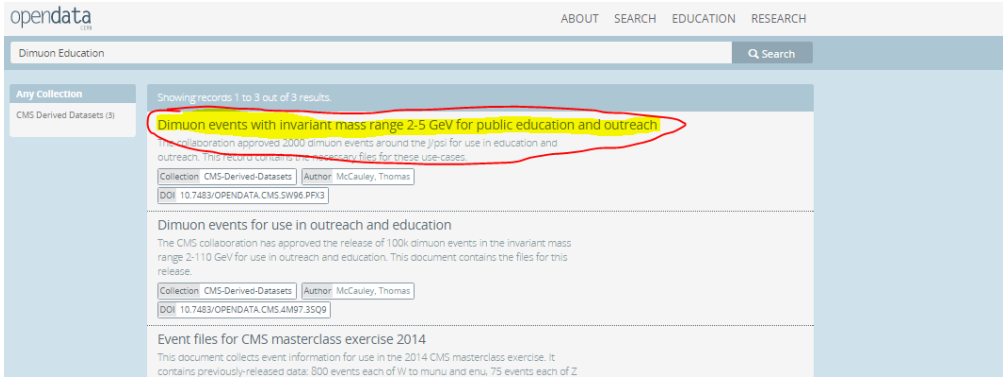
Click "Start Learning" in Education section.

Click "Explore CMS".

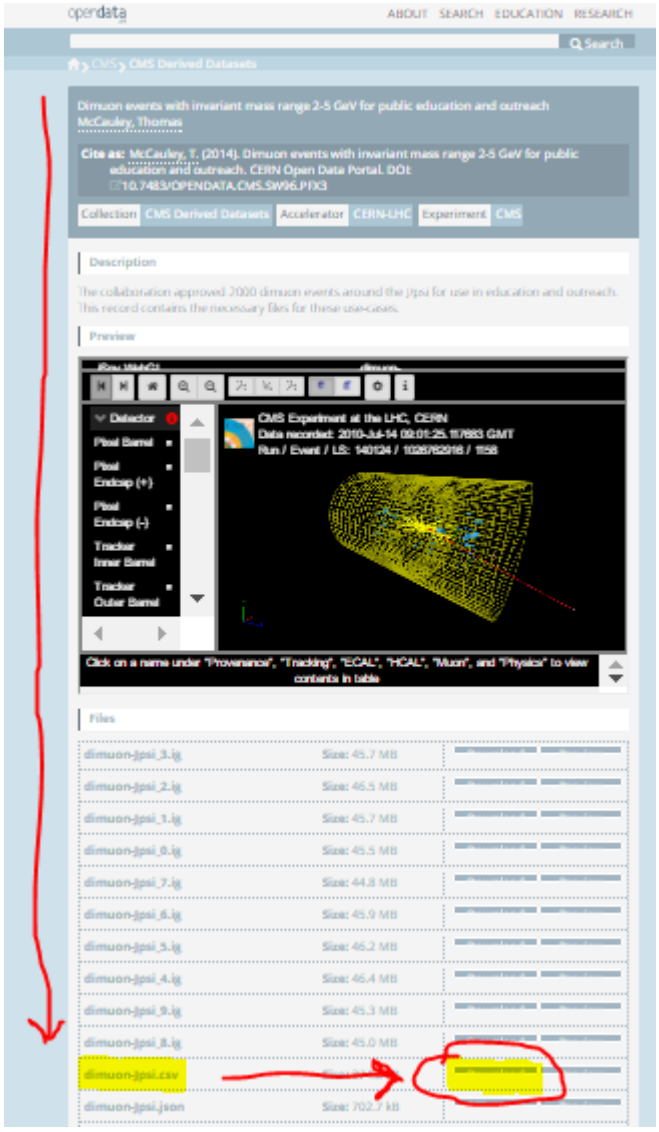
Type "Dimuon Education" into the search bar at the top of the page and click "Search"



Click "Dimuon events with invariant mass range 2-5GeV for public education and outreach".



Scroll down and go to “dimuon-Jpsi.csv” and click “Download”:

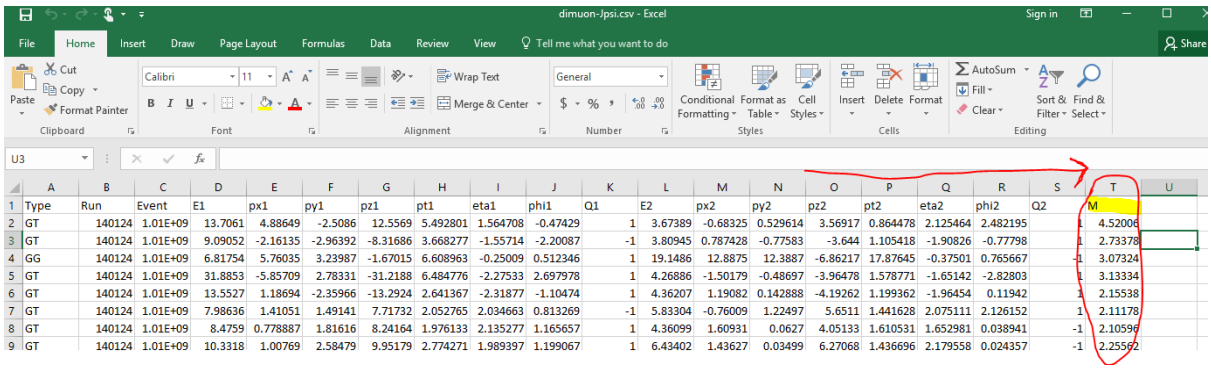


Open this file in using Microsoft EXCEL.

2.) Create a Histogram of Invariant Mass

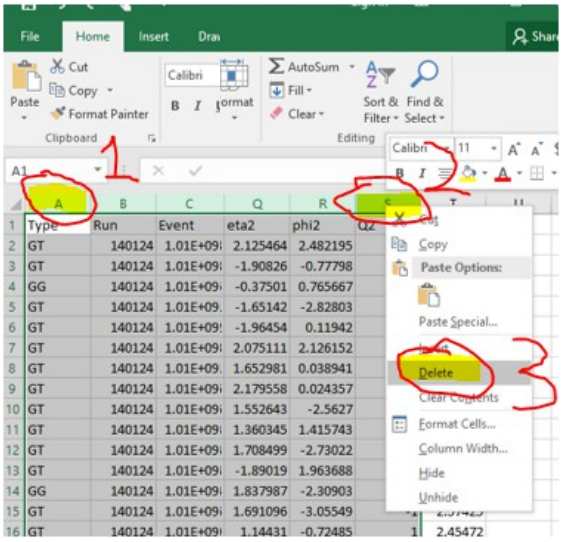
2.1) Prepare Data

In the Excel file there is a lot of data. Each row corresponds to an event where two muons were detected. Each column corresponds to a piece of data from that event. For this task we only care about the final column “M”. “M” stands for invariant mass and is measured in GeV (Giga Electron Volts).

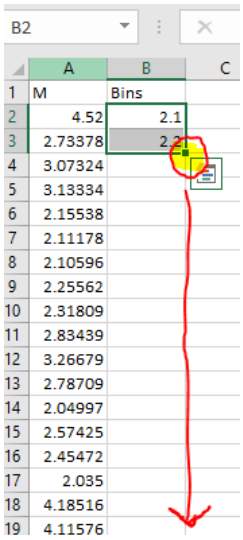


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Type	Run	Event	E1	px1	py1	pz1	pt1	eta1	phi1	Q1	E2	px2	py2	pz2	pt2	eta2	phi2	Q2	M	
2	GT	140124	1.01E+09	13.7061	4.88649	-2.5086	12.5569	5.492801	1.564708	-0.47429	1	3.67389	-0.68325	0.529614	3.56917	0.864478	2.125464	2.482195	1	4.52006	
3	GT	140124	1.01E+09	9.09052	-2.16135	-2.96392	-8.31686	3.668277	-1.55714	-2.20087	-1	3.80945	0.787428	-0.77583	-3.644	1.105418	-1.90826	-0.77798	1	2.73378	
4	GG	140124	1.01E+09	6.81754	5.76035	3.23987	-1.67015	6.608963	-0.25009	0.512346	1	19.1486	12.8875	12.3887	-6.86217	17.87645	-0.37501	0.765667	-1	3.07324	
5	GT	140124	1.01E+09	31.8853	-5.85709	2.78331	-31.2188	6.484776	-2.27533	2.697978	1	4.26886	-1.50179	-0.48697	-3.96478	1.578771	-1.65142	-2.82803	1	3.13334	
6	GT	140124	1.01E+09	13.5527	1.18694	-2.35966	-13.2924	2.641367	-2.31877	-1.10474	1	4.36207	1.19082	0.142888	-4.19262	1.199362	-1.96454	0.11942	1	2.15538	
7	GT	140124	1.01E+09	7.98636	1.41051	1.49141	7.71732	2.052765	2.034663	0.813269	-1	5.83304	-0.76009	1.22497	5.6511	1.441628	2.075111	2.126152	1	2.11178	
8	GT	140124	1.01E+09	8.4759	0.778887	1.81616	8.24164	1.976133	2.135277	1.165657	1	4.36099	1.60931	0.0627	4.05133	1.610531	1.652981	0.038941	-1	2.10596	
9	GT	140124	1.01E+09	10.3318	1.00769	2.58479	9.95179	2.774271	1.989397	1.199067	1	6.43402	1.43627	0.03499	6.27068	1.436696	2.179558	0.024357	-1	2.25562	

Delete all the other columns (A to S). To do this select all the columns by clicking on “A” **(1)**, then holding down the ‘shift’ key and then clicking “S” **(2)**. Then right click on “S” **(2)** and click “Delete” **(3)**.



In cell B1 (next to the M column) write a new column called “Bins”. In cell B2 type “2.1” In cell 3 type “2.2”. Click the little green square (see diagram) and drag it down to row 31. This should create a list of values from 2.1 to 5. It should look like the image shown on the right (below) when finished.



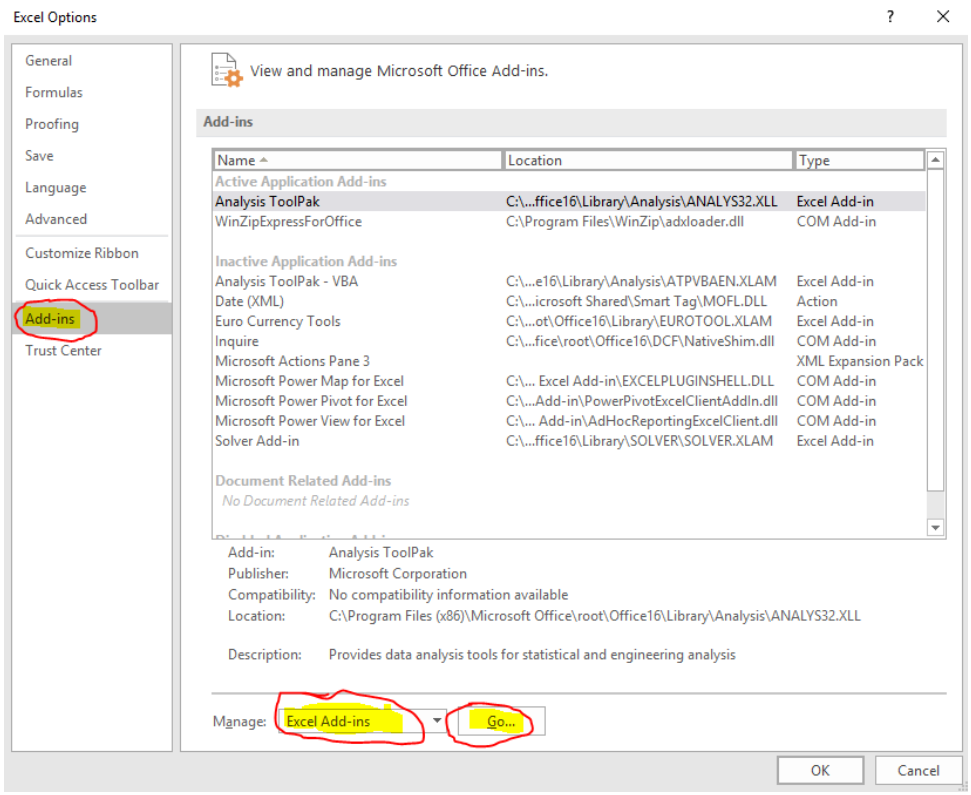
	A	B
1	M	Bins
2	4.52	2.1
3	2.73378	2.2
4	3.07324	2.3
5	3.13334	2.4
6	2.15538	2.5
7	2.11178	2.6
8	2.10596	2.7
9	2.25562	2.8
10	2.31809	2.9
11	2.83439	3
12	3.26679	3.1
13	2.78709	3.2
14	2.04997	3.3
15	2.57425	3.4
16	2.45472	3.5
17	2.035	3.6
18	4.18516	3.7
19	4.11576	3.8
20	3.61339	3.9
21	2.56937	4
22	3.7036	4.1
23	2.08016	4.2
24	2.51015	4.3
25	2.10543	4.4
26	2.60879	4.5
27	2.5664	4.6
28	3.4788	4.7
29	3.20743	4.8
30	2.5911	4.9
31	3.0094	5

2.2) Active Excel Histogram package

Click “File” on the top left-hand side of the screen.

Click “Options” tab on the lower left hand side of screen.

Click “Add-ins” tab on the lower left hand side of the window. Then select “Excel Add-ins” and click “Go”.

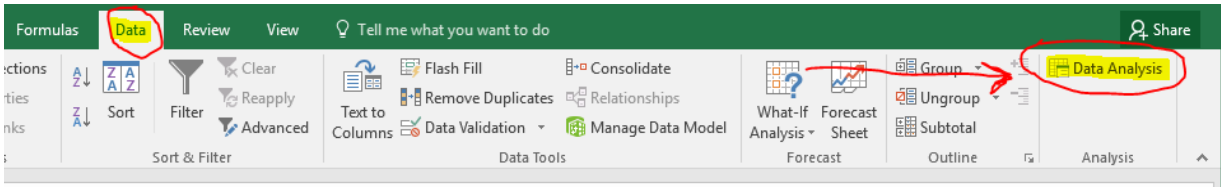


Tick the Analysis ToolPak box and click OK.

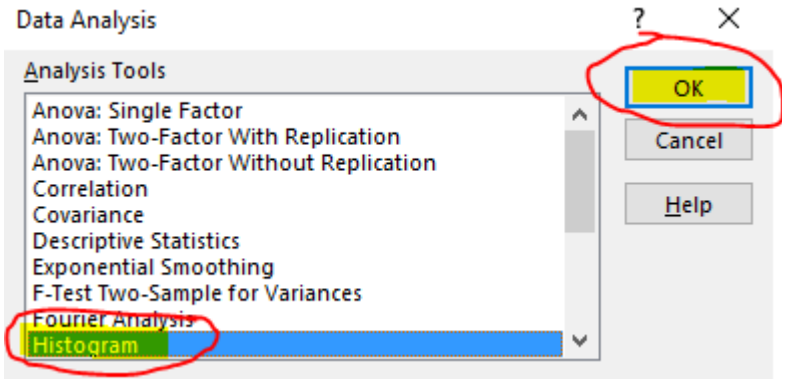
2.3) Prepare Data for Histogram

Click on the data tab.

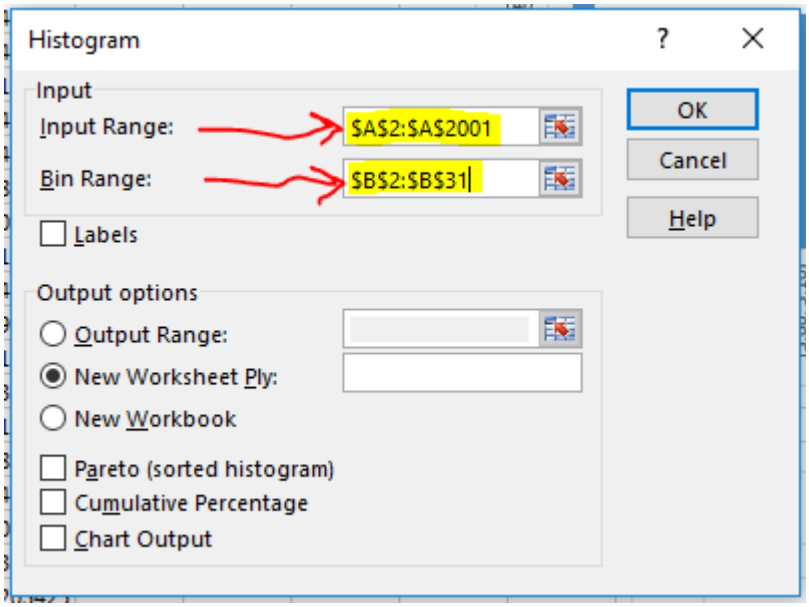
“Data Analysis on the right hand side”.



Select “Histogram” and click “OK”.

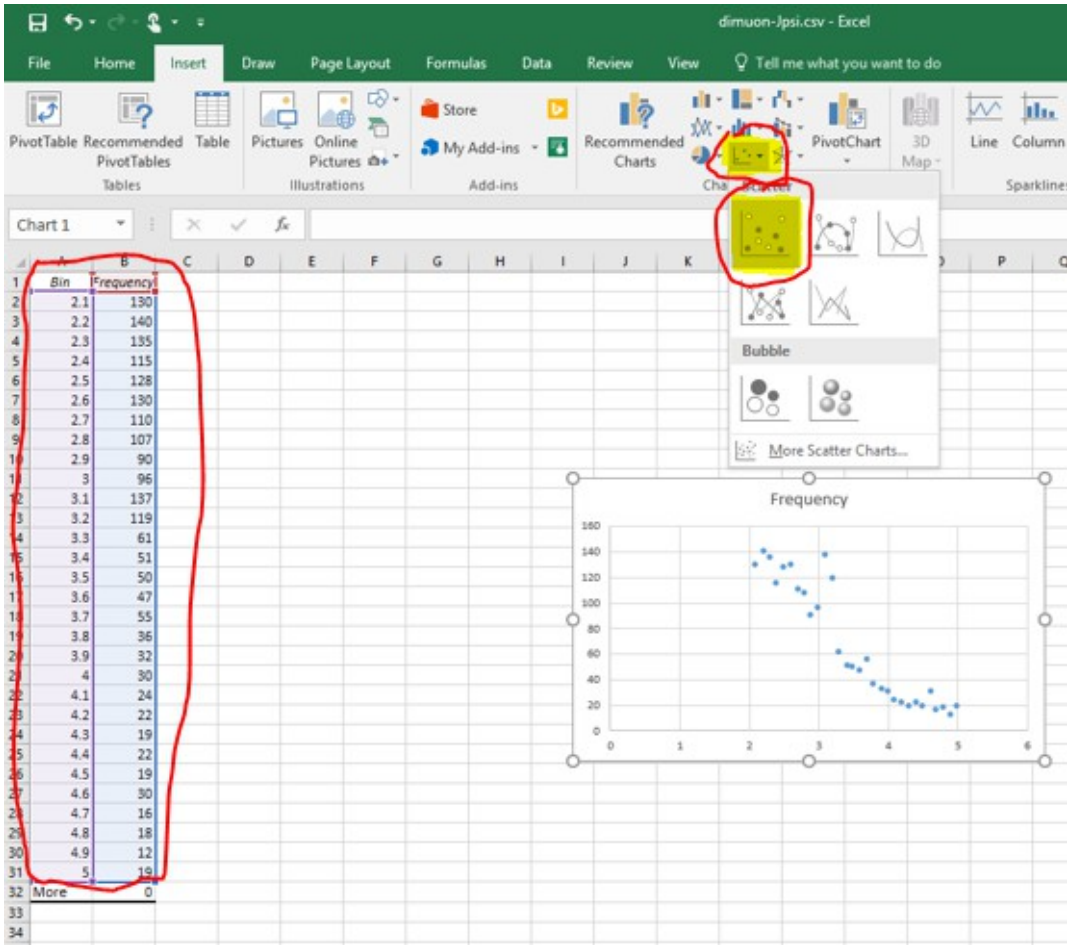


In the “Input Range” type “$A$2:$A$2001”. In “Bin Range” type “$B$2:$B$31”. Click OK.



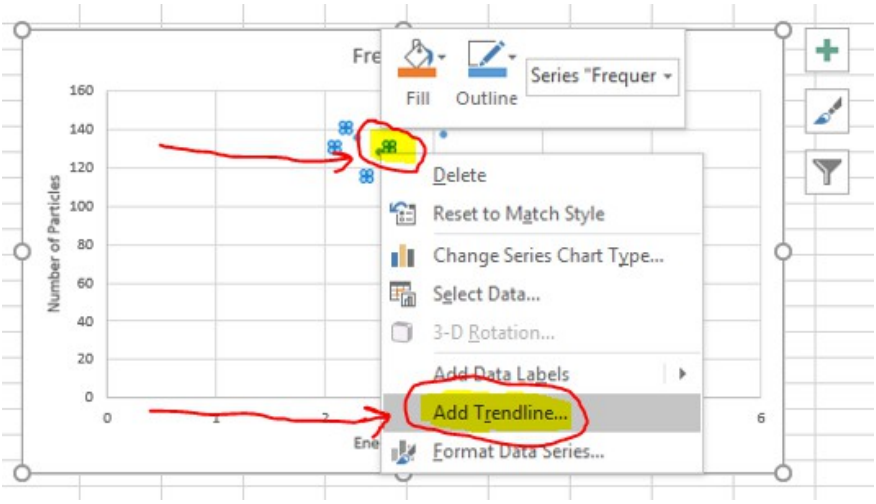
2.4) Create Histogram (Scatterplot)

Go to the newly created sheet, you can access it by clicking on “Sheet 1” tab on the bottom left hand side of the screen. Select cells A1 to B32. Click the “Insert” tab, click on the scatter icon and then click the “Scatter” diagram on the top left corner.

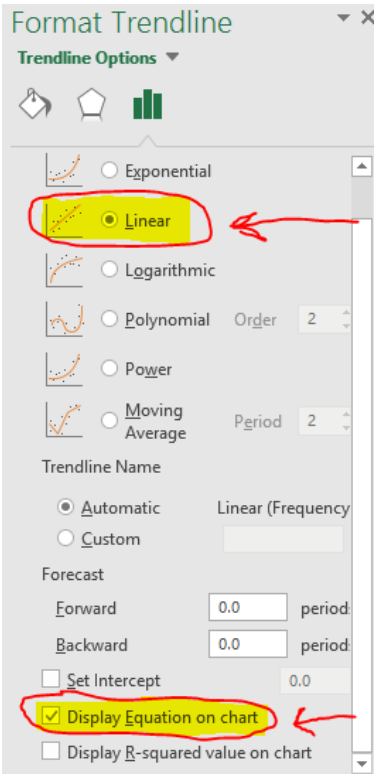


Add units on the axis labels (Y-axis is number of particles; X-axis is Energy of Particles (GeV)).

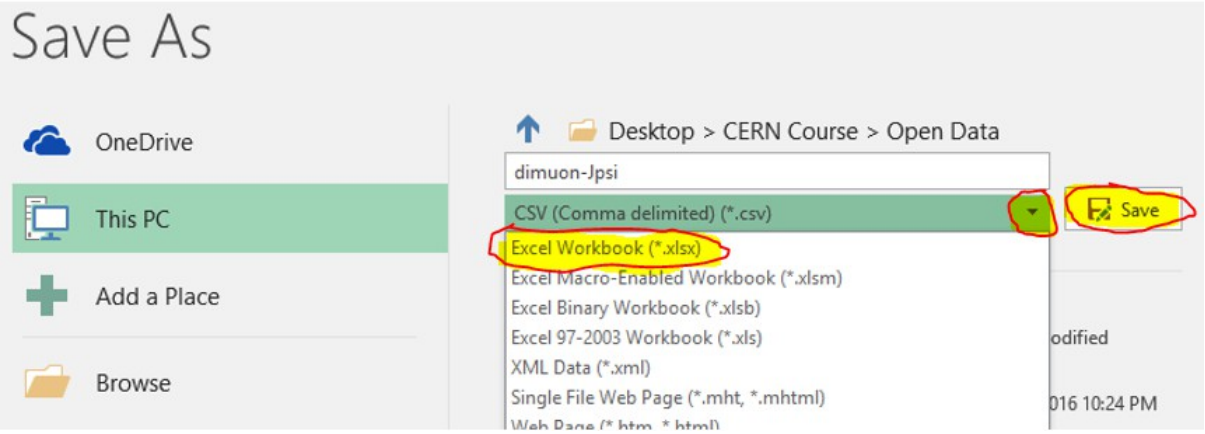
Add a trendline to the data, to do this first click on a data point. Then right click the data point then click “Add Trendline”. You can do this by selecting the “Design” tab and clicking “Add Chart Element” on the left hand side of the screen.



Then select “Linear” and then tick the “Displace Equation on chart” box:



Save your workbook as an .xlsx file. To do this click “file”, then click “save as” on the left hand side of the screen. Click the “This PC icon”, then click the down arrow next to the save button and select “Excel Workbook (*.xlsx)”. Then click th save icon.



2.5) Questions

a) Write down the equation for your line of best fit.

b) What is the gradient of your line? Is the gradient positive or negative?

c) What is the y-intercept? How many particles would you expect to detect at 0 GeV?

d) Does the fit seem to be a good fit to the data, are there any sections which look unusual?

3.) Plot Graph by Hand

Get a sheet of graph paper and try plotting the data from the histogram. Now draw in a trendline by hand. The line should aproximately fit the data and approximately half the points should be above the line and half below.

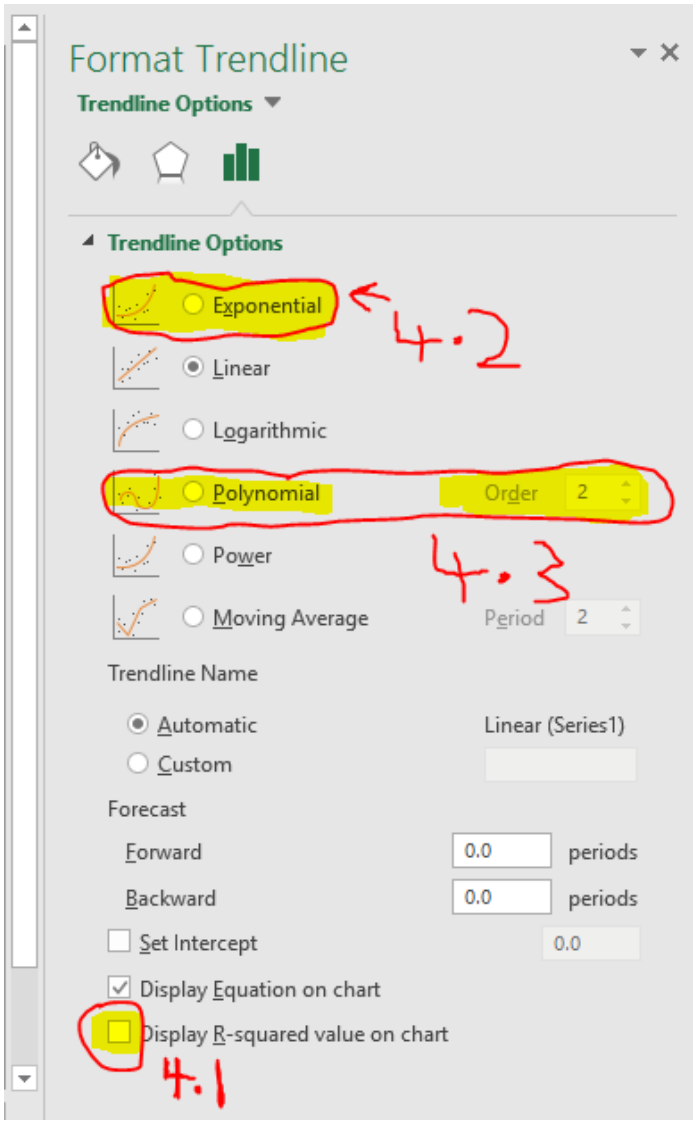
a) What is the gradient of your line? Is the gradient positive or negative?

b) What is the y-intercept? How many particles would you expect to detect at 0 GeV?

c) Does the fit seem to be a good fit to the data, are there any sections which look unusual? Is there any evidence of a “bump” which could be a particle?

4.) Apply non-linear fit to the data

Go to the histogram completed at the end of section 2. Click on the trendline, then right-click on the trendline and select “Format Trendline”.



4.1) Create a measure of Goodness of fit

In format trendline there are a number of options available. These allow you to use different types of equtions to fit the data. Tick the “Displace R-squared value on chart” box. This will cause a number called R^2 to appear below your equation on the chart. The closer this number is to one, the better the fit is to the data. However R^2 is not always a good measure of the quality of the fit.

a) What is the R^2 value for your linear fit?

4.2) Expotential Fit

Click “Exponential”. Write down the R^2 value.

a) Is the R^2 value bigger or smaller than that of the linear fit?

b) Does expoential look like a good fit to the data? What are the problems with it?

4.3) Polynomial Fit

Click “Polynomial”. Write down the equation and R^2 value for the fit.

a) Is the R^2 value bigger or smaller than that of the linear fit?

b) Does the parabola look like a good fit to the data? What are the problems with it?

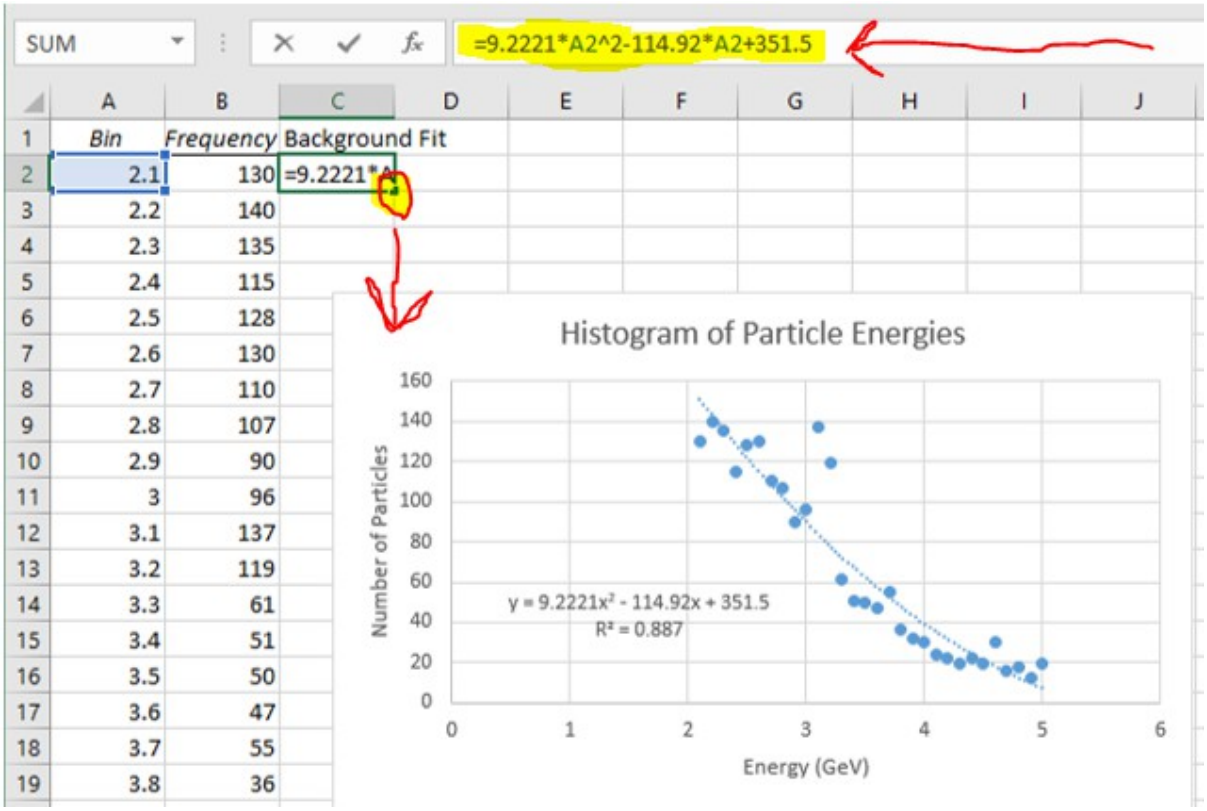
Note it is possible to increase the order of the polynomial for the polynomial fit by clicking the up and down arrows on the order button. Increasing the order will always give a higher R2 value, however this does not necessarily mean the fit is better, the polynomial may just be copying the shape of the noise (error) rather that finding the trend of the data.

5.) Subtract off Background

5.1) Create Background Values

To complete this section, you will need the data from the previous sections and you will need the equation of the parabola (order 2 polynomial) from the previous section.

Into cell C1 type “Background Fit”. Then in cell C2 type “=” and then your parabola equation. Then press “Enter”. Note in excel to type the times symbol use the “*” key (shift+8). To raise something to a power use the “^” key (shift+6), for example 3^2 would be written as “3^2”. Note that your x-value will be the number in cell A2, you can simply select the cell rather than type it if you wish.

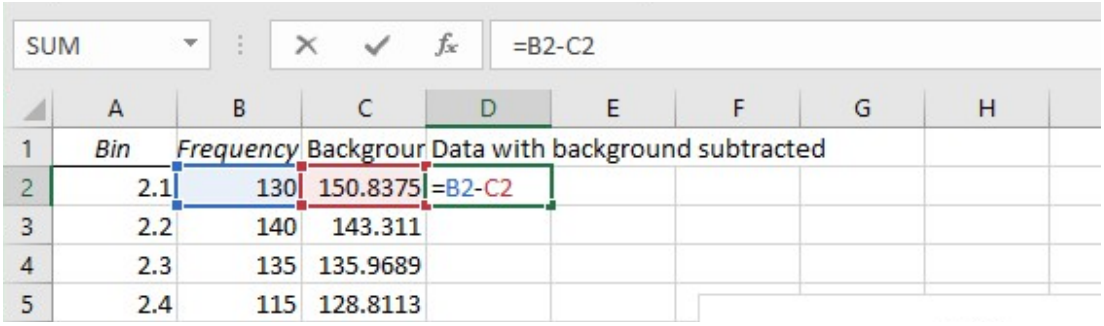


	A	B	C	D	E	F	G	H	I	J
1	*Bin*	*Frequency*	Background Fit							
2	2.1	130	=9.2221*A							
3	2.2	140								
4	2.3	135								
5	2.4	115								
6	2.5	128								
7	2.6	130								
8	2.7	110								
9	2.8	107								
10	2.9	90								
11	3	96								
12	3.1	137								
13	3.2	119								
14	3.3	61								
15	3.4	51								
16	3.5	50								
17	3.6	47								
18	3.7	55								
19	3.8	36								

Click on cell “C2”. Click and hold the green box in the bottom right corner of the cell and drag it down.

5.2) Subtract Background from the data

Click on cell “D1” and type “Data with background Subtracted”. In Cell “D2” type “=B2-C2”.

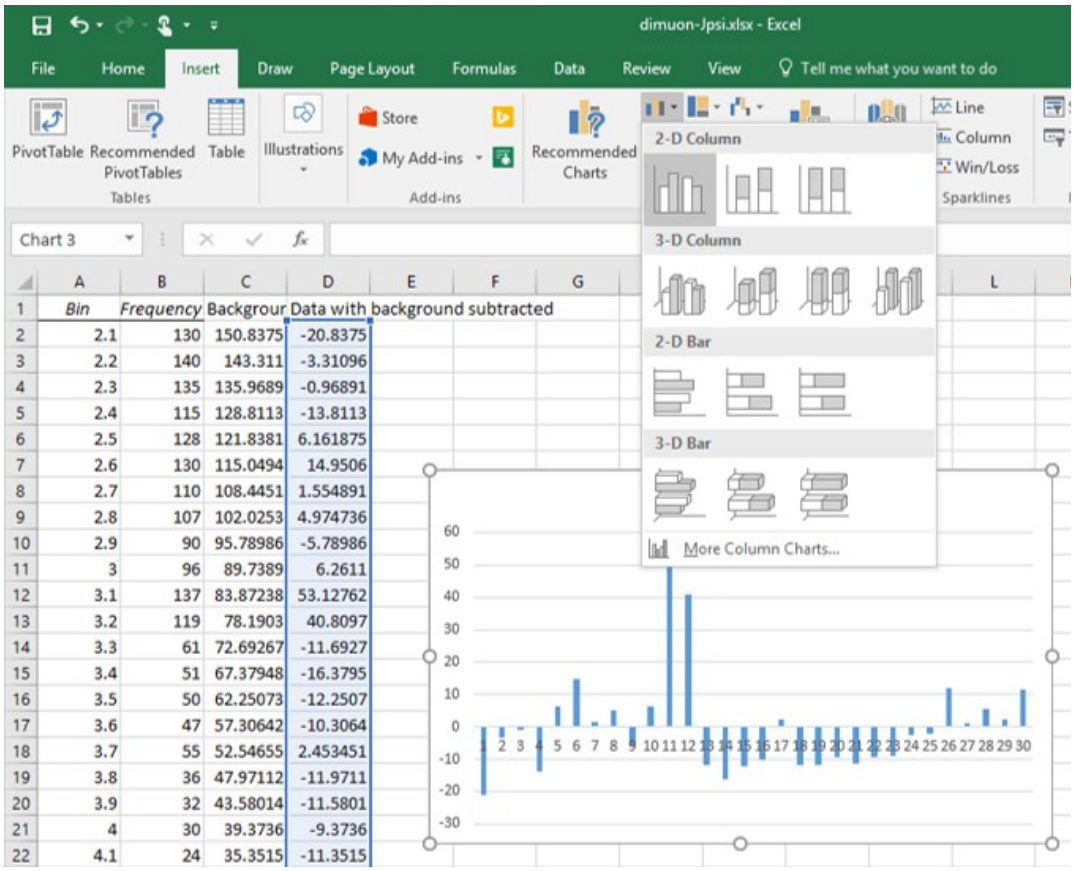


	A	B	C	D	E	F	G	H
1	*Bin*	*Frequency*	Backgrour	Data with background subtracted				
2	2.1	130	150.8375	=B2-C2				
3	2.2	140	143.311					
4	2.3	135	135.9689					
5	2.4	115	128.8113					

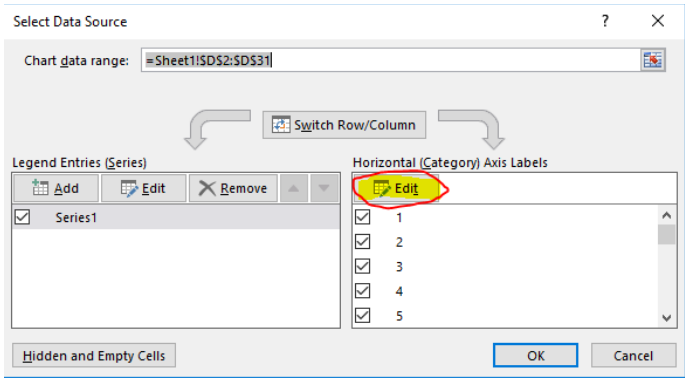
Drag the green box on the bottom right hand of the screen to copy the equation down for each row.

5.3) Plot the Data with the Background removed

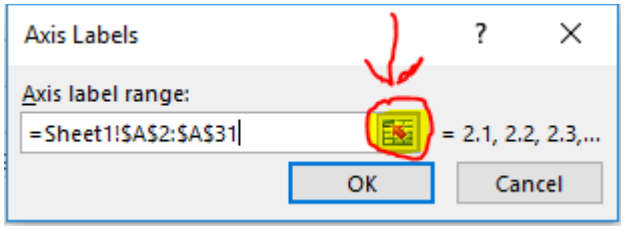
Select cells “D2” to “D31”. Then click insert tab at top of page and click arrow next to “Column Chart” icon. Then click 2-D Clustered Column.



Click on the numbers below the x-axis. Then right-click one of the numbers and click “Select Data”. Click “Edit” (Horizontal Axis Labels).

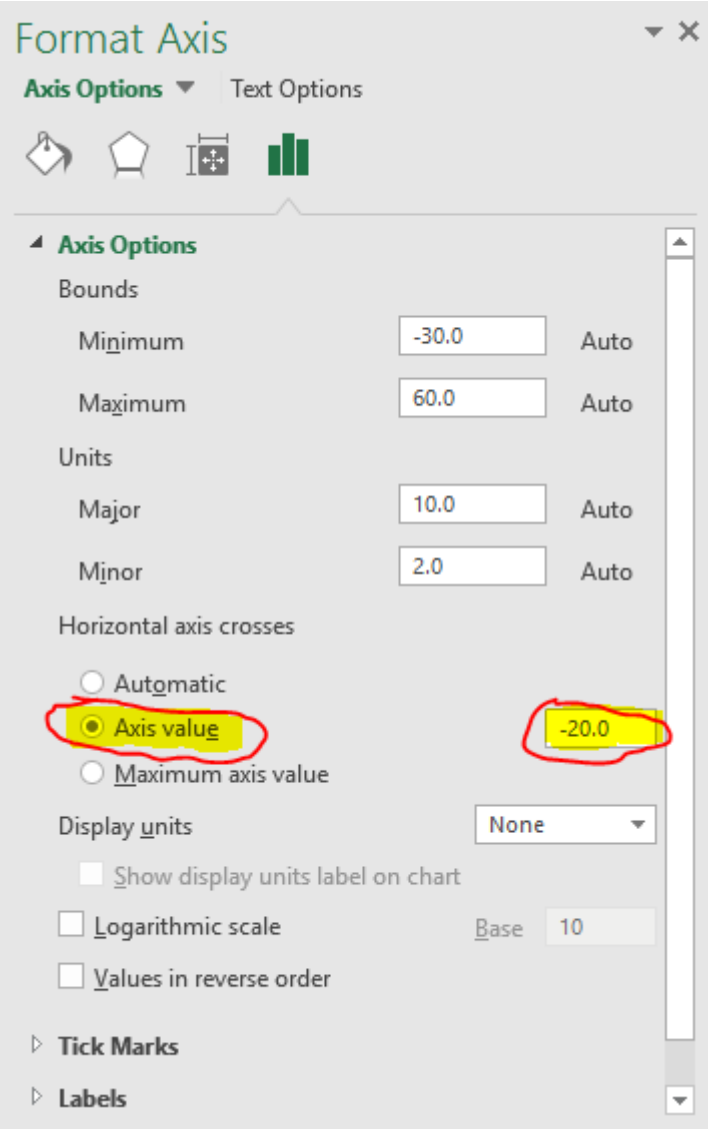


Click the small square with the arrow and select cells “A2” to “A31”. Then click OK.



Add a chart title and axis labels. You can do this by selecting the "Design" tab and clicking "Add Chart Element" on the left hand side of the screen.

Click on the y-axis on the left hand side of the chart. Right click and select "Format axis". Click the "Axis value" and put in the value "-20" .



5.4) Examine the Plot

a) Are there any columns which seem significantly higher than the others? At what energy does this occur?

b) How much higher are these columns that the others?

c) Do you think there is a new particle? Give evidence for you conclusion.

6.) Test for statistical significance

In the previous section you drew a conclusion about whether the data indicates a new type of particle. In this section we will go through a more rigorous statistical process that scientists use. There are two pieces of terminology you will need to know in order to do this:

- Mean – The mean of a set of data is the same thing as the average. The indicates the approximate size of the data. It is calculated by adding up all the values and dividing by the number of values.
- Standard Deviation – The standard deviation is a measure of how spread out the data is. If there is a high standard deviation, then the spread in the data is large. If there is a low standard deviation, then the spread in the data is low.

When looking at the data it appears that there may be a peak at 3.1 and 3.2 GeV. This corresponds to the energy of a J/Psi Meson

6.1) Find the mean and standard deviation of "bump"

Set up cells in rows G to J ready for calculations. Simply type the relevant words into the cells as shown below:

	A	B	C	D	E	F	G	H	I	J
1	*Bin*	*Frequency*	Backgrour	Data with background subtracted						
2	2.1	130	150.8375	-20.8375						
3	2.2	140	143.311	-3.31096						
4	2.3	135	135.9689	-0.96891				Mean	Standard Deviation	
5	2.4	115	128.8113	-13.8113			3.1 and 3.2 GeV:			
6	2.5	128	121.8381	6.161875			All other data:			
7	2.6	130	115.0494	14.9506						
8	2.7	110	108.4451	1.554891			Combined Standard Deviation:			
9	2.8	107	102.0253	4.974736			Number of Standard Deviations:			
10	2.9	90	95.78986	-5.78986						

In cell H5 calculate the mean of 3.1 and 3.2 GeV. You can do this by typing the following formula "" into cell H5. Calculate the standard deviation of 3.1 and 3.2 GeV using this formula "".

6.2) Find the mean and standard deviation of all other data.

Now calculate the mean and standard deviation for all other energies by modifying the formulas shown in the previous section. You can select the relevant cells using "D14:D31,D2:D11" .

6.3) Combine standard deviations

In order to find the total standard deviation, you need to type this formula "=SQRT(I6^2+I5^2)" into cell "I8".

6.4) Count how many standard combined deviations there are between the means.

In cell “I9” calculate the difference between the means and divide this by the combined standard deviation.

a) How many standard deviations are there between the means?

In order to prove you have discovered a new particle in physics you need to show that your results have a certain statistical significance. If the number of standard deviations between the means is less than 5, then this is not considered statistically significant and you cannot rigorously conclude that the particle exists.

b) Would physicists consider this result to be significant?

When you have finished the above instructions:

Try changing the size of the bins in the histogram and using different orders of polynomials to fit the data. Are you able to make your result significant?