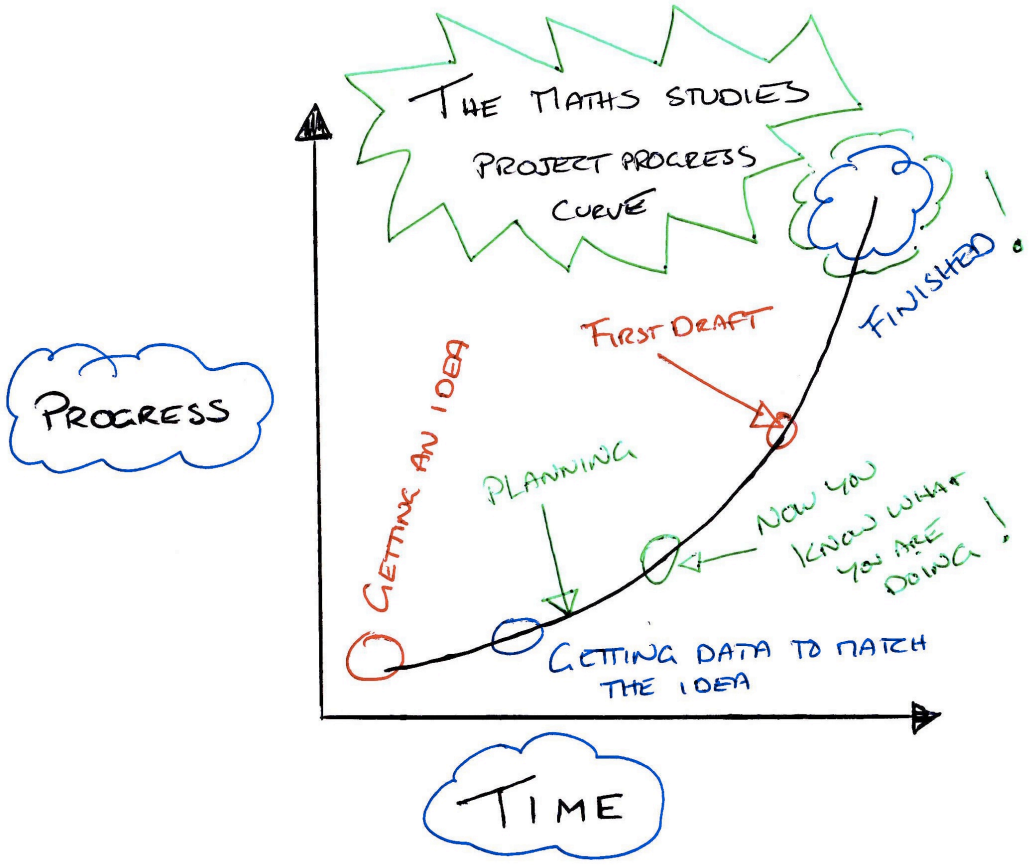
Mathematical Studies

Internal Assessment Project

Writing Guide



Internal Assessment Guide

This guide is written with a particular focus on statistically based projects. Read the guidance carefully and pay attention to the details. The guide will not do your project for you, but should help you make your project clear and coherent.

This guide assumes that you have been through the process of choosing an idea and are now ready to carry the project out and write about it.

The Introduction – Title, Statement of task and Plan

General Advice

- This is your opportunity to set the tone for the whole project. Be clear, make sure there is a title, then there should be an explanation of what you hope to achieve and why you thought it might be interesting. Then you should give a detailed plan of the steps you are going to take. Be as specific as possible.
- You probably need to revisit your introduction after you have completed the project work. You might not even write it fully before then.

Score	Description
0	There is no clear statement about what the project is about – No title or anything that tells the reader about what is coming
1	There is a clear statement about what the project is about – It can be easily understood from the beginning of the project what it is about.
2	The above is done and with this there is a clear title and a description of the plan. The plan would be a summary of exactly what you intend to do – for example, how you are going to gather information and what things you are going to do with it. It helps to be as specific as possible in your plan here.
3	All of the above is done *and* you follow the plan . This means the reader must see everything you said you were going to do. Your plan should be specific. I.e. 'I am going to do a scatter graph of this against this……'

What is your project about?

It is excellent practice to be able to describe easily and clearly the nature of your project. In general, for statistics based projects, you must be able to answer the question 'What are you surveying?'

Examples of 'What are you surveying?'

I am surveying 200 ***people***
I have collected information about 160 different ***cars***
I am surveying the top 100 ***movies*** of all time as voted by IMDB users
I have data on all of the world's 204 different ***countries***

More answers……

Songs, youtube channels, books, planes, cities, images, musicals……

Now answer the question for your project 'What are you surveying?'

Titles

It may seem obvious, but this is a key element that it is important to get right because the project must stick to the title and the plan.

Consider the following example of a project done about 'People' who had their body mass index (BMI) measured….

'Investigating factors that may influence a persons BMI'

or

'Is a persons BMI dependent on Gender?'

Both titles may be appropriate, but the second one limits you to a specific line of inquiry about Gender where the first is more open. Again, both could be good projects, it is just about making sure you choose a title that best matches your project.

Have a go at writing your title now

Statement of task

Again, this is seemingly straightforward, but you need to make sure you have explicitly stated this. Pursuing the example above about BMI, you should say something like…

'I am going to collect information from 200 people and measure their BMI. I will then look to see which factors may have an influence on BMI. For example, I will look to see if BMI might be dependent on gender, age or nationality.'

This is a statement of task and can be elaborated on in the plan.

Try to write a statement of task about your project

The Plan

This is a crucial element and often makes the difference between a 2 and a 3 in this criteria.

- Your plan should be as detailed as you can make it and should include everything that follows.
- For this reason, you will want to check the plan again at the end to be sure that it matches.
- To get the top mark here, you need to explain the purpose of the processes you are planning to use.

Consider the following example (an extract from a plan)

I will then plot a scatter graph of film budget against the worldwide gross of the top 100 domestic grossing films.

Or

I will then plot a scatter graph of film budget against the worldwide gross of the top 100 domestic grossing films. I will be looking to see whether there is a correlation between budget and the worldwide gross. For example, is it true that 'The more a movie cost to make, the more money it will take at the box office'?

The first simply states what will happen, the second says what the purpose of doing this is. If you are not sure what the purpose is, you probably shouldn't be doing it!

Try and write one part of your plan with its purpose. You shouldn't really go any further until your plan is clear

The Information

General Advice

- If you have collected primary data then you should include a copy of the questionnaire/data collection tool you used
- If you used secondary data then you should explain how you decided which data to collect and use. This is called the sampling process
- Try to tell the whole story of the information gathering process. What did you do? Where did you go looking? Who did you ask? How did you decide on the data you used? This is often the longest part of a project and so there is usually plenty to tell.
- WHAT are you surveying? You should be able to answer this question! It might be people, cars, shapes, countries, songs, etc etc. and then you have some information about each of them. The amount you need does depend a little on what it is you are collecting info about, but 50 (people or cars or songs etc.) is a good minimum. DO consult your teacher on this
- If it is more of an maths investigation than a statistical analysis then there is more to think about, but it can be done.

Score	Description
0	There is no actual information or data and no evidence of any attempt to collect any.
1	There is some relevant data collected This means that it is related to your stated aims. Even if there is a problem with the data or the way in which it was collected, you can still get this mark.
2	Your data is well organized so that you can do what you wanted to do with it. This probably means that it is in a table that is clearly understood. It is a good idea to have all of your information in ONE table. (see above on 'what are you surveying) OR You have enough (see above) good information. Good means that it will be suitable for you to carry out your processes. It also means that there isn't a problem with it and the way it was collected. You also need to have told the story of your data collection to get this one.
3	All of the above! So AND not OR

You should have a section in your write up that details the information you have collected, the process you went through to collect it and a justification of your decisions.

Description

Start by describing the data you have, then it is a good idea to include a small sample of your data to help the explanation. The rest can then be included as an appendix. See the example below…..

RANK	FILM	Studio	Budget ($)	Domestic Gross ($)	Worldwide gross ($)	YEAR	Genre	IMDb rating
1	Avatar	Fox	425 000 000	760 509 625	2 783 918 982	2 009	Sci Fi	8.0
2	Titanic	Par	200 000 000	658 672 302	2 185 672 302	1 997	Drama	7.7
3	Marvel's The Avengers	BV	225 000 000	623 357 910	1 514 279 548	2 012	Action	6.6
4	The Dark Knight	WB	185 000 000	534 858 444	1 002 891 358	2 008	Action	9.0
5	Star Wars: Episode I - The Phantom Menace	Fox	115 000 000	474 544 677	1 007 044 677	1 999	Sci Fi	6.5
6	Star Wars: Episode IV - A New Hope	Fox	11 000 000	460 998 007	797 900 000	1 977	Sci Fi	8.8
7	The Dark Knight Rises	WB	275 000 000	448 139 099	1 079 343 943	2 012	Action	8.6
8	Shrek 2	DW	70 000 000	441 226 247	937 008 132	2 004	Animation	7.3
9	E.T: The Extra-Terrestrial	Uni	10 500 000	435 110 554	792 965 326	1 982	Sci Fi	7.9
10	Pirates of the Caribbean: Dead Man's Chest	BV	225 000 000	423 315 812	1 060 615 612	2 006	Fantasy	7.3

The database contains information on the top 100 movies as ranked by IMDB. For each Movie, I have collected

- *The studio the movie was made in*
- *The budget,*
- *The domestic gross (money taken in ticket sales)*
- *The year the movie was made*
- *The IMDB user rating*

I then added a field for 'Genre' so that I can compare different types of movie.

Story telling

It is then important to describe the process you went through to get the data. This might include…..

- Talking about the different sources you found and explaining why you chose the one you used.
- Talking about the process of designing a questionnaire and getting it filled in.
- Describing the process of setting up an experiment designed to collect data. (eg setting up the gym to find out what weight people can bench press)
- Describing problems you encountered along the way
- Talking about anything you added to help split the data in to categories (Like the example above with 'genre')

Primary data - considerations

Primary data is that which you have collected by using a questionnaire or conducting an experiment. If you are using primary data you should consider the following….

- Make sure you talk through the whole process of designing a questionnaire or experiment.
- What questions did you ask and why?
- What measurements did you take and why?
- What inconsistencies might have arisen from the way you collected data?
- How much information did you collect? What factors affected this?
- Include your questionnaire or data collection sheets as an appendix. Can you include photographs of the experiment?

Secondary data – considerations

Secondary data is that which you have obtained from external sources who have compiled the data already. If you are using secondary data you should consider the following….

- Make sure you have described and referenced your sources
- Consider whether or not sampling is appropriate (should you use all the data you found or take a sample of some of it)
- Explain why you used all OR only some of the data you found
- What sampling has already been done by your source? Eg - your source only has data on countries in Africa OR the source only listed top 100 songs of all time.

Amount of Data

It is important to justify the amount of information you have. This essentially means explaining why you didn't use more or less. The following might be possible reasons…..

- *This was all the data my source had*
- *I would not have been practical to have collected more because of the time involved*
- *Taking a sample of this data would only have reduced the accuracy of my findings so it was more appropriate to use it all.*

Organisation of data

It is important that your data appears all in one organized place as an appendix to the project. In most cases it is highly recommended that the data is organized in ONE spreadsheet that can be printed to be included. (If you don't include this you can only get 1/3 here).

You might consider describing how you have grouped or ordered the data. This might be in Rank order, or by continent or genre for example.

The Mathematics

General Advice

- RELEVANCE is the key here for getting more than 2 marks. It is easy to miss something here so do check carefully. A common mistake might be something like the difference between 'Number of crimes' and 'crime rate'. This is an important difference and the one you use might make a difference to the relevance of the process you use.
- Watch out for things like doing a scatter graph and an independence test with same data. The latter will be considered irrelevant.
- Further processes might be – calculus, modeling, linear regression, correlation coefficient, independence tests, distributions or compound probability.
- A further process needs to involve step-by-step calculations that you show with full working out. For example – if you do an independence test you need to show how the expected frequencies are calculated and then how you have worked with them to calculate the chi squared statistic. If you did more than 1 test then you only need to show the calculations for 1.
- Get the balance right! Less but correct and relevant is better than more but incorrect and irrelevant.

Score	Description
0	There is no maths! You haven't done any calculations with your information.
1	You have done at least 2 simple bits of maths. This might be calculating an average, standard deviation, drawing box plots or scatter graphs and more. Check with your teacher whether or not a process you are doing will count as a simple process. Anything that you do that uses *only* technology counts here as simple.
2	The above is done and correctly. You might get away with a couple of mistakes but be careful. It is more important to do less, but correctly, than more with more mistakes.
3	The above, but now that **ALL** (not just two) of the processes are correct and relevant (see above).
4	All of the above are true AND you have tried a relevant 'further process' (see above) even if this contains mistakes.
5	The above, but the further processes are correct. You need at least 1 further process. This also depends on the quality of your information. If it is not good enough then you might not qualify for this mark.

Simple and further

The following are just examples and the list is not exhaustive.

Examples of simple processes

- Mean, median, mode,
- Range and Interquartile range,
- Standard deviation
- Box plots and cumulative frequency
- Pie charts, bar graphs and histograms
- Scattergraphs, correlation and linear regression done using technology only

Examples of further processes

Note – for these to be considered, many elements of the calculations need to be shown. If they are done using technology only, they will not count.

- Calculation of correlation coefficient
- Calculation of regression line
- Chi² test of independence

Correctness

This is a key point. For credit to be given, calculations should be largely without error. Less calculations that are correct are worth more than more that contain errors.

Where possible, you should confirm calculations using technology and show that you have done so.

Relevance

It is important to make sure that the processes that are used are relevant. This does take some careful consideration. The following are a list of typical irrelevance…

- A bar chart and pie chart to show the same data – unless you can demonstrate that one offers something that the other doesn't, one of these might be considered irrelevant.
- The inclusion of several calculations (often from technology) that you do not use or talk about.
- A line of best fit on a scattergraph that does not how correlation. You should check first if there is sufficient correlation ($r < -0.5$ or $r > 0.5$) before you plot, talk about or use a regression line.
- A regression line that you do not use.
- ANY PROCESSES that do not help to answer your question or investigate your theme.

There are other more subtle examples of irrelevance. It is good practice to try and explain (in a sentence) why the process you are using is relevant.

Like with correctness, it is often better to use less processes but make sure they are relevant.

The Simple Processes

This section shows good and bad examples of these processes being used

Measures of central tendency (mean, mode, median, range sd, iqr)

It is always a good idea to consider some simple data analysis like those listed above. These are often good descriptors and help both the writer and reader 'get a feel' for the data set.

See these examples…..

This table needs a heading so that it is clear. eg *'The mean and standard deviation for the number of views for photographs in diofferent genres'*

	Mean	Standard Deviation
Lifestyle	85120.68	159367.66
Architecture	28647.3	50163.7
Portrait	57389.3	75320.52
Landscape	106078.3	182747.32
Macro	188995.63	281116.97
Wildlife	24383.08	40202.07

DANGER - Having made the calculations, it is important to use them in your discussion that answers your question - other wise it risks being seen as irrelevant

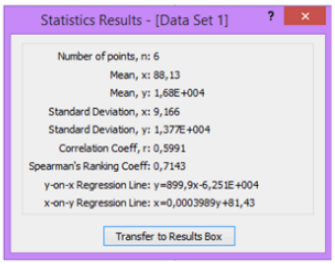
Tables such as these are useful summaries of the data analysis and make it easy to see all in one go. Discussion can then follow easily.

Average books published per year		Average literacy rates	
Africa	810	Africa	68,9
Asia	21339	Asia	89,03
Europe	27300	Europe	98,23
North America	39240	North America	91,53
Australasia	3181	Australasia	88,525
South America	8943	South America	92,575

BEWARE - Consistency with notation is important -

What rounding has taken place?
Is the comma a decimal point?

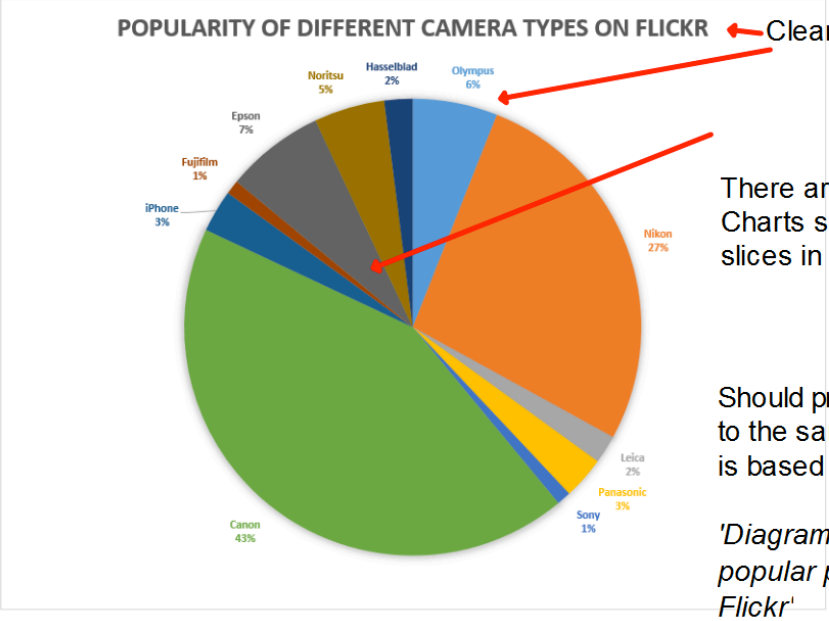
Make sure you are consistent and clear about what you are doing.



TECHNOLOGY - often students include screen shots of calculations done with technology. This is fine, but make sure you say so and point out which of those calculations you are going to use.

Pie Charts and Bar graphs

See this example of a pie chart



Clear title and labels

There are those that feel Pie Charts should always have the slices in order of size

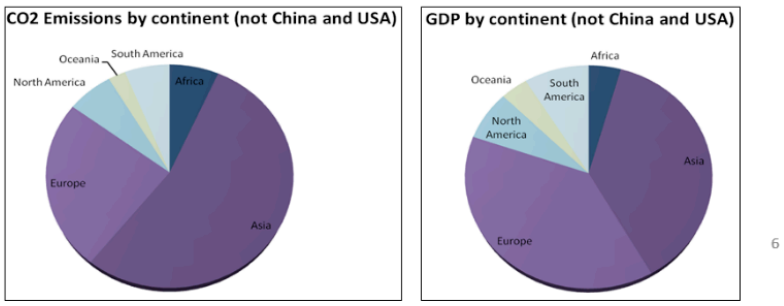
Should probably make a reference to the sample on which this diagram is based

'Diagram based on 100 most popular photographs on Flickr'

The interpretation should probably appear beneath the diagram, but could easily be added with annotations. This can make it easier to understand.

See this example

The pie charts for annual CO2 emissions and for annual GDP by continent are as follows, again without China and the United States:



6

From this we see that, although Europe and Asia (omitting data from China) have very similar GDP, the CO2 emissions are rather different. According to this data, Asia emits more than twice the amount of CO2 than Europe does. Therefore, compared to their GDP, Asia and Europe's CO2 emissions show complete inequality.

The Camparison is usually a good idea. Can you make conclusions from looking at this diagram?

What is NOT clear is that the pie shows the 'Average GDP for countries in that continent' NOT the GDP of the continent (eg)

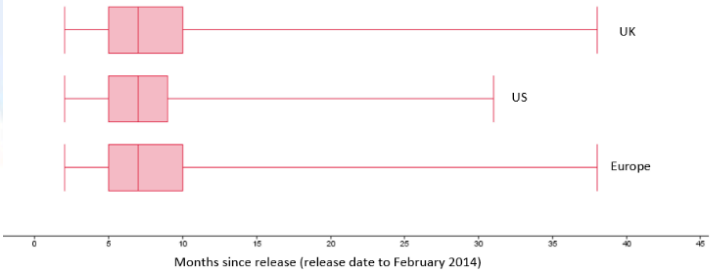
Box plots

Box plots or bow and whisker diagrams are a really nice way to compare 1 variable data distributions. Check out the examples below about how these can be used and what considerations you should make.

- Do make sure you have enough data to make a box plot. One with 3 data items for example would not be valid. If in doubt, ask.

Box and whisker graph: Shows the lower quartile, upper quartile, median and the range of the months since the release of the songs for the UK, US and Europe.

What songs?



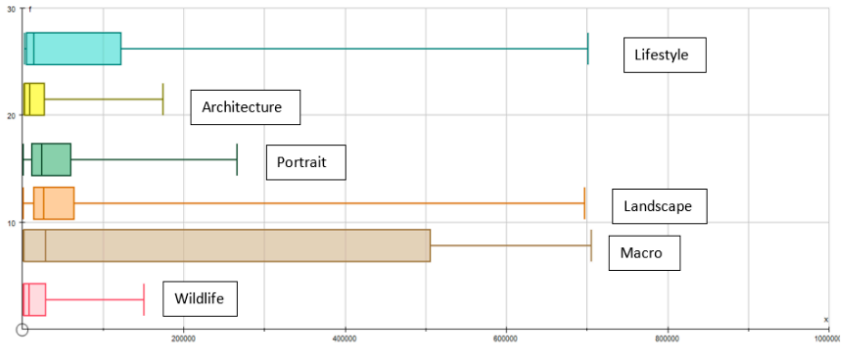
Box plots should be showed on the same axis so that they can be easily compared

Each plot should be labelled to aid the comparison

Be aware that the technology you use may include a vertical axis that has no meaning. You should point this out.

Be really clear in describing what the box plots show.

1. **Amount of views per photo for each genre**



The above box and whisker diagram shows the total amount of views that each photo has accumulated within each genre from the 100 Flickr users that I surveyed.

Is this clear enough?

Be aware of the effect of 'outliers'. One result can stretch the extremes and the scales.

If you remove outliers you should explain and could include before and after diagrams

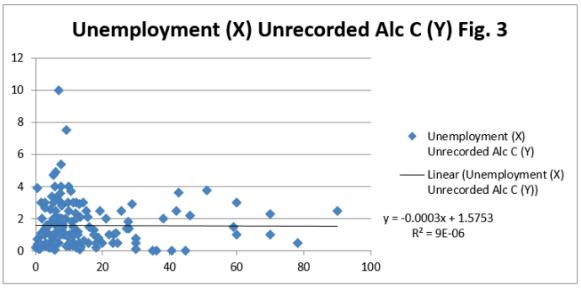
Make sure you label the horizontal axis and consider the scale you use. In your intepretation you should point out and try and explain both the similarities and differences.

Scattergraphs, correlation, regression

The fine details of these are underestimated. Useful tools but there are lots of things to consider……. The following is a list of key rules to try and follow.

- Scattergraphs are for comparing two numerical data fields
- Make sure that there is a title and that the axes are labelled
- Make sure the purpose is clear and clearly stated
- Make a visual judgment about whether or not there is correlation. Confirm this by quoting 'r' the correlation coefficient. (Remember R^2 is not the same as r.)
- Only include a line of best fit or regression line when there is sufficient correlation.
- Use a reference to help you decide what kind of correlation you have.
- Make sure you don't use computer notation to state either the value of 'r' or the equation of the regression line. (Beware software that will do this automatically)
- IF you quote a regression line, THEN, you should use it to make a prediction within the range of the data you have.
- IF you want to use it to make a conjecture about the future, THEN fine, but explain that you know it is not valid because it is out of the range.

See the examples below for one or two examples of things that can go wrong with scatter graphs.



Title is not clear, axes are not labelled

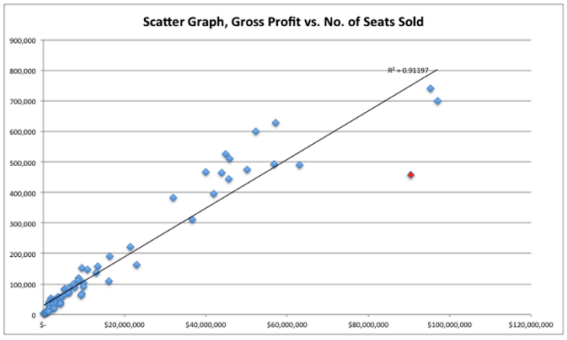
Points are bunched together because the scale is not big enough

There is no obvious correlation and yet a line of best fit is drawn.

R^2 confirms no correlation

R^2 is not the correlation coefficient, 'r' is.

Excel uses computer notation for standard form - not acceptable



This diagram is better, BUT, the student still referred to R^2 as the correlation coefficient in the interpreation.

Axes are not labelled

The Further processes

Correlation coefficient and Regression lines

If you want calculating the Regression line or the Pearson's Product Moment Correlation Coefficient to be considered as a further process then consider the following.

- You should explain the whole process as you go.
- You should construct a spreadsheet that shows all the stages of the calculation. (See example below)
- You must be correct and consistent with all the notation and terminology. (eg, don't write 'xbar' for the mean)
- You MUST show how you substitute the different sums in to the formula.
- Use technology to confirm your answer.

See the example below for calculating the PMCC

When the two sets of data x and y are the two sets of data, the formula for the Pearson Product-Moment Correlation Coefficient is: $r = \frac{s_{xy}}{s_x s_y}$

where s_{xy} is the covariance and s_x and s_y are the standard deviations of x and y.

$$s_{xy} = \sum \frac{(x-\bar{x})(y-\bar{y})}{n} \qquad s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} \quad s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

Above the Brandt Line:

Table 3: A sample of the table that finds the covariance of firearm-related deaths per 100,000 people and guns legally owned per 100,000 people above the Brandt Line. The full table can be found in Appendix C.

Country (Above Brandt Line)	Value (x) (Gun Deaths per 100,000 people)	Value (y) (Guns Legally Owned per 100,000 people)	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})(y - \bar{y}) \div n$ $n = 43$
South Africa	21.51	12700	19.04976744	-6600	-125728.4651	-2923.917793
United States	10.3	89000	7.839767442	69700	546431.7907	12707.71606
Montenegro	8.55	23100	6.089767442	3800	23141.11628	538.1654949
Serbia	3.9	58200	1.439767442	38900	56006.95349	1302.48729
Switzerland	3.84	45700	1.379767442	26400	36425.86047	847.1130341
...	...	...	...	...	...	...
Ukraine	0.2	6600	-2.260232558	-12700	28704.95349	667.5570579
Romania	0.19	700	-2.270232558	-18600	42226.32558	982.0075717
Singapore	0.16	500	-2.300232558	-18800	43244.37209	1005.683072
Belarus	0.1	7300	-2.360232558	-12000	28322.7907	658.6695511
Japan	0.06	600	-2.400232558	-18700	44884.34884	1043.822066
					Σ (covariance) :	20682.53488

Table 4: A sample of the table that finds the standard deviations of firearm-related deaths per 100,000 people and guns legally owned per 100,000 people above the Brandt Line. The full table can be found in Appendix D.

Country (Above Brandt Line)	Value (x) (Gun Deaths per 100,000 people)	Value (y) (Guns Legally Owned per 100,000 people)	$(x-\bar{x})^2$	$(y-\bar{y})^2$
South Africa	21.51	12700	362.8936396	43560000
United States	10.3	89000	61.46195354	4858090000
Montenegro	8.55	23100	37.0852675	14440000
Serbia	3.9	58200	2.072930287	1513210000
Switzerland	3.84	45700	1.903758194	696960000
...	...	...	...	...
Ukraine	0.2	6600	5.108651217	161290000
Romania	0.19	700	5.153955868	345960000
Singapore	0.16	500	5.291069822	353440000
Belarus	0.1	7300	5.570697729	144000000
Japan	0.06	600	5.761116333	349690000
		$\sum(x-\bar{x})^2$:	533.2224977	$\sum(y-\bar{y})^2$: 12530100000
		$\sum(x-\bar{x})^2 \div 43$:	12.4005232	$\sum(y-\bar{y})^2 \div 43$: 291397674.4
		$\sqrt{\sum(x-\bar{x})^2 \div 43}$: (Standard Deviation)	3.521437661	$\sqrt{\sum(y-\bar{y})^2 \div 43}$: (Standard Deviation) 17070.37417

$$r = \frac{\text{Covariance}}{\text{Standard Deviation of x} \times \text{Standard Deviation of y}} \qquad r = \frac{20682.53488}{3.521437661 \times 17070.37417} = 0.344065$$

Independence tests

These are good to include and can provide a nice development of the project. Read the bullet points about how to run these tests and then see the examples.

- Chi² Tests should involve looking for a relationship between data fields where one or more of them are categorical.
- Make sure the purpose is clear and related to the question/theme.
- Think about the limits of a contingency table. 2 x 2 is probably too small, 6 x 6 probably too big. Obviously it depends on the context of the problem. If in doubt ask.
- Make sure your contingency tables are correctly compiled and labelled.
- If you are categorizing a numerical data field then you MUST explain and justify the method you used to create categories.
- State clearly the null and alternate hypotheses.
- Display a table of expected frequencies.
- If you have a 2 x 2 table then you must use 'Yates's continuity correction'
- Demonstrate how at least 2 of these expected frequencies were calculated. You can do them all.
- You must not have any expected frequencies less than 5. If this happens you need to consider merging categories. You must justify how you combined categories. Remember the point about 2 x 2 tables above.
- Construct a table that shows all the stages of the calculation resulting in the chi² calculated statistic. (Spreadsheets are good for this)
- Confirm this result using technology.
- Demonstrate how the 'degrees of freedom' are calculated.
- Decide on a significance level – you might consider giving the result 3 times for each of 1, 5 and 10%.
- Use a table of Chi² critical values (and reference it) to arrive at the critical values needed for the test.
- Accept one of the hypotheses.

Yates's continuity correction - If you use this, you need to talk about it

Categories	FO	FE	FO-FE-0.5	(FO-FE-0.5)²	(FO-FE-0.5)²/FE
Male yes	1	7,666667	-7,1666667	51,36111111	6,699275362
Male no	22	15,33333	6,16666667	38,02777778	2,480072464
Female yes	19	12,33333	6,16666667	38,02777778	3,083333333
Female no	18	24,66667	-7,1666667	51,36111111	2,082207207
					14,34488837

See the following example – there is no need to follow it exactly (subheadings etc) and it could have been done differently, but it does contain all the key elements listed above.

Chi² Tests of Independence, Step-By-Step Calculation and Analysis

My first Chi² Test will be to examine whether the gender of a YouTuber and the amount of subscriptions are independent of each other. In order to explain how I got to my conclusion my calculations will be set out in a step-by-step format.

Step 1 – Write a null hypothesis (H□) and an alternative hypothesis (H□)

Essentially, the null hypothesis states that the two sets of data are independent and the alternative hypothesis states they are not independent, or dependent.

H□: The amount of subscriptions on a YouTube channel is independent of the gender of the YouTuber.

H□: The amount of subscriptions on a YouTube channel is not independent of the gender of the YouTuber.

Step 2 – Classify your data into specific groups if numerical

Obviously if you are looking at gender, the data is categorical and so does not need to be categorised. However, I am working with enormous numerical data, subscribers to YouTube channels, and so the data needs to be organised. My groups are as follows:

- ***Low amount of subscribers = 0 – 2,600,000***
- ***Medium amount of subscribers = 2,600,001 – 4,000,000***
- ***High amount of subscribers = 4,000,001 +***

My intervals are not equal so that the distribution of the data is more even across the Chi² Test. This will eradicate the possibility of the expected frequency values being less than five.

Step 3 - Note down your observed frequencies

The observed frequencies are the frequencies that you have witnessed. It is often useful to put them in a contingency table.

	Low	Medium	High	Totals
Female	9	7	15	31
Male	19	48	40	107
Mixed	8	18	18	44
Totals	36	73	73	182

Step 4 – Calculate the expected frequencies

Calculating the expected frequencies can be done by using the formula for the probability of independent events.

For example:

The expected number of female YouTuber's subscribers being in the low category would be =

31/182 x 36 = 6.131 (3.d.p)

The expected number of mixed YouTuber's subscribers being in the high category would be =

44/182 x 73 = 17.648 (3.d.p)

	Low	Medium	High	Totals
Female	6,1319	12,4341	12,4341	31
Male	21,1648	42,9176	42,9176	107
Mixed	8,7033	17,6484	17,6484	44
Totals	36	73	73	182

Step 5 – Calculate the Chi² Statistic

When working out the Chi² Statistic by hand the following formula must be used[1].

$$\text{Chi}^2 = \Sigma \frac{(\text{Observed frequencies} - \text{Expected frequencies})^2}{\text{Expected frequencies}}$$

$$= \Sigma \frac{(Fo - Fe)^2}{Fe}$$

This is more easily seen in a table like the one below, displaying my results:

	Fo	Fe	Fo-Fe	(Fo-Fe)²	(Fo-Fe)²/Fe
Female Low	9	6,131868	2,868132	8,22618	1,34154555
Female Medium	7	12,43407	-5,43407	29,52907	2,3748525
Female High	15	12,43407	2,565934	6,584018	0,52951445
Male Low	19	21,16484	-2,16484	4,686511	0,22142914
Male Medium	48	42,91758	5,082418	25,83097	0,6018738
Male High	40	42,91758	-2,91758	8,512287	0,19834032
Mixed Low	8	8,703297	-0,7033	0,494626	0,05683206
Mixed Medium	18	17,64835	0,351648	0,123657	0,00700669
Mixed High	18	17,64835	0,351648	0,123657	0,00700669
Chi² Statistic	5,338401213				

Step 6 – Note the level of significance

The most common levels of significance are 1%, 5% and 10%.

For my Chi² Test I have chosen to use a level of significance 5% or 0.95.

Step 7 – Calculate the number of degrees of freedom

To find the number of degrees of freedom you must use a formula based on the contingency table:

Degrees of Freedom = (number of rows – 1)(number of columns – 1)

Degrees of Freedom for Gender/Subscribers Chi² Test = (3-1)(3-1)

= 2 x 2

= 4

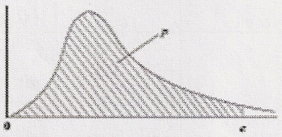
[1] http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage28.htm

Step 8 – Work out the critical value

Using my degrees of freedom at 4 and a level of significance of 0.95, we can tell from the table below that my critical value is 9.488.

Critical values of the χ^2 distribution (topic 6.9)

$p = P(X \leq c)$



p	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
$\nu = 1$	0.00004	0.0002	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

ν = number of degrees of freedom

6

Step 9 – Compare your calculated Chi² Statistic against the critical value

If your Chi² Statistic is less than the critical value then you do not reject the null hypothesis.
If your Chi² Statistic is more than the critical value then you reject the null hypothesis.

Step 10 – State your conclusion

Since the Chi² Statistic is less than the critical value, I will reject the alternative hypothesis and accept the null hypothesis. Therefore, leaving us with the conclusion that the amount of subscriptions on a YouTube channel is independent of the gender of the YouTuber.

Interpretation

General Advice

- For each bit of mathematics you do you should get a result and state it clearly.
- You should try and explain what that result means.
- You should try and draw inferences from the results ie – why do you think they might be that way
- If you can link some of your results together then you are doing really well.
- Remember, the aim of the project was to investigate something you are really interested in to see if you could find out anything. Did you?

Score	Description
0	You have not made any conclusions or interpretations based on your processes.
1	You have made at least one interpretation or conclusion, even if it is wrong.
2	You have made interpretations and conclusions that are correct based on your processes. (this means that you can get this mark even if your processes are wrong or flawed) . There is no 'magic number here' but as a rule you should make a statement of conclusion to go with each of your processes – otherwise, why do it?
3	You write really well and thoroughly about the results you have from the maths you have done. This is a hard one to get. Try to make sure you have followed a pattern of stating the result, explaining what that means and making an inference for why that might be the result.

It ought to be easy to get to 2/3 here, but 3/3 can be quite demanding. In many ways, this depends on the nature of the project and the results that came from it.

Some guidelines for interpreting as well as you can….

- Make sure you have some mathematical processes to interpret.
- Make sure your interpretation is consistent with the result of the process.
- Relate your interpretation to the question/s or theme you stated at the beginning in your introduction.
- Talk about the mathematical significance of the result. (This mean
- Talk about the contextual significance of the result.
- Try to explain/infer reasons that may be behind the results.
- Try and cross reference different results from different parts of the exercise.

Validity

General Advice

- Key – it is not enough just to say that should have collected more data you need to say why and what you think the impact of more data would have been. For example, you might suggest that because you only asked IB students your results cant be considered representative of the whole population because 18 yr olds will have different habits to 40 year olds.
- The validity of processes is also important and has an impact on relevance in criterion C. For example, when is it appropriate to show a regression line? You could get this mark for commenting on the validity of processes.
- If you think that there are no issues with validity then you have to explain why. In practice this is seldom the case.

Score	Description
0	You have not mentioned validity at all or shown any awareness that it is an issue
1	You have specifically mentioned the validity of your investigation or a process within it and explained why you think it is an issue.

Must have mark!

Validity is always an issue and there are many ways to show how you have considered this. When you give an example of 'validity' as an issue, you should explain 'why' you think this is an issue and the effect it has on the outcome.

The following are some examples….

'Too many of the people in my survey were in the same age range. This limits the validity of my conclusions because the other age ranges were not properly represented. To improve on this I would need get more responses from people in other age groups'

'I had a numerical data field in my chi² test and had to divide the data in to categories to continue. I have explained how I did this, but I could have done it other ways and that might have had an effect on the outcome. As such, the method for choosing categories affects the validity of the outcome'

Structure and Communication

General Advice

- Signposting is really important. Good use of headings, sub headings and page breaks can make it really easy for the reader. They want to know where they are and what is happening at any given moment.
- Reading well is a subjective judgment BUT easily tested. Give your project to someone unconnected with the course to read. How many times do they stop to ask you a question? One that reads really well will be easy for a third party to follow, even if they don understand all of the maths.
- Again, less is more – less discussion that is meaningful and relevant scores better than more discussion that rambles and loses focus.
- Be careful with titles and labels for graphs and diagrams etc. These could easily come under this criterion and the last. You shouldn't get penalized twice, but guard against it by being clear.

Score	Description
0	There is no structure ion the project at all.
1	You have made some attempt to structure the project. This might apply to really simple, short or unfinished projects.
2	You have structured the project in a logical order so that it is easy to follow. There is an implication that this is complete project that has all the key elements.
3	The project does the above, it reflects the stated plan AND is well written so that it reads well. Footnotes and bibliographies help effective communication and contribute to getting this mark. Again – the discussions you have must be relevant.

Points to consider

MUST DO - Get someone else to read your project. The questions they ask you tell you everything you need. It is easy to miss things when you are so involved in the work. You know what it is about and how it all fits together. The only way to find out if it makes sense to a reader who has not been involved is to test it out.

Structure

- Let your structure match up with the introduction. This way, if your plan is logical, then your project will be too.
- You should be able to describe the sections of your project in conversation – this helps to make it clear in your mind.
-
- Use a clear 'Headings system and sub headings system' (Word does this very well)
- Try to start new key sections on a new page.
- Avoid new sub headings very close to the end of a page. (The page break interrupts the flow)
- Reference as you go using 'footnotes'
- Then include the references you use in a bibliography at the end. (There are plenty of online bibliography tools to help with this)
- Use the same font and size throughout.

Graphs/Charts/analysis

- Make sure graphs are titled and labelled correctly and clearly. It is very frustrating when you have to look for these details.
- Always helpful to annotate charts and graphs to point out the key features.
- Discussions for analysis should follow straight afterwards and – if possible – on the same page.

Notation and Terminology

General Advice

- Give graphs and diagrams clear titles and labels
- Define the variables that you use
- Use the correct terms for the processes you use (see examples from your classwork/books to check)
- Watch out for computer notation use 'X' instead of '*' for multiply etc. Use 3.4 x 10^{-6} instead of 3.4E-6 or 10^-6
- A little care is all that is need to get the best marks here.

Score	Description
0	There is no correct notation and terminology
1	There is some correct notation OR terminology. The implication is that there are errors with one or the other.
2	The notation and terminology is correct all the way through. If the project is too simple then you cant really get this one because it doesn't require enough notation and terminology.

Important!

The following are some examples of things to avoid/watch out for.

- You need to have actually done some maths in order to get credit here.
- When using letters to stand for numbers, state explicitly what quantity each letter is standing for.
- Look at examples from different sources, to make sure you are using the correct terminology – eg Null and alternate hypotheses for chi² tests. Scatter graphs can show 'correlation' chi² tests show 'dependence'
- If you are showing long calculations for correlation coefficient or regression, show $\bar{x}$ as the mean, not 'xbar'
- Don't use '*' for multiply or 3E2 for 3 x 10² (standard form)

If in doubt ask!

As a rule 1 mistake here can be overlooked if there enough other good examples. 2 mistakes will probably stop you getting 2.