1.0 Introduction:

I have always had an interest in Statistical studies. I found it satisfying to synthesis and analyse information to draw meaning conclusion out of seemingly nothing. The idea for my exploration came to me when a relative of mine from China reached out to me in a dilemma. She is a bright student, and had been a top scorer for the subject she takes. However, she is currently taking three humanities and she needed to choose one subject to drop. She wanted to know if there was some ways to measure exactly how well she had performed in comparison to each subject.

At the same time, during my IB course, I learned about normal distribution which includes standard deviation and also z-scores. This tells me that I should be able to determine a statistical measure to compare her performance of a subject to that of another. Different class of students have different range of capabilities, therefore scoring a same mark doesn't mean the student is performed equally in all three subjects. From what I've learnt about normal distribution, since a normal distribution is dependent on the data's mean and standard deviation, it suggests to me that, in different data sets, having the same raw test scores does not tell me how it compares to other students in the same test.

In other words, a student with the same mark may have different z-scores which allows us to show contrast for the student's performance in each subject. That is why for my exploration I aim to compare performance of a student in different subjects.

2.0 Review of materials:

2.1 Normal distribution

To better understand how z-scores are compared, we shall first take a closer look at the normal distribution. Normal distribution is a probability function that describes the tendency for how a set of data would be distributed. It is a symmetrical distribution where most of the data values observed would cluster around the central peak, the mean value, and the probabilities for data values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are therefore similarly unlikely. Data that follows a normal distribution is defined by two parameters, its mean and standard deviation. [1]

[2]

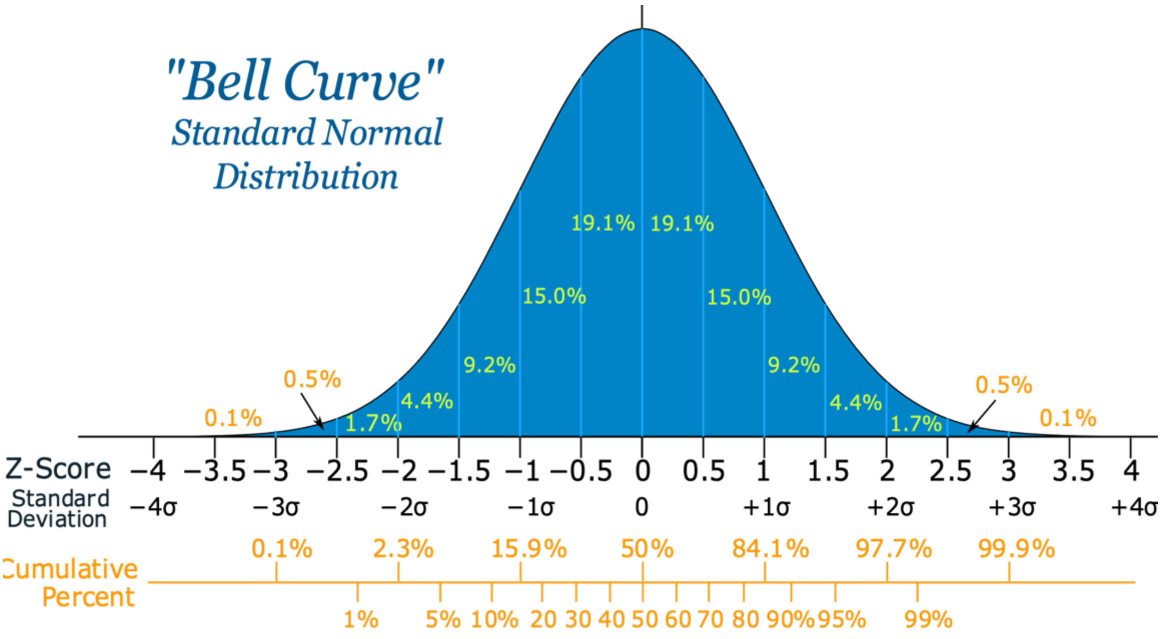


Diagram 1: Normal Distribution

[1] Russell, D. (2019, September 3). What Is a Bell Curve, Anyway? Definition and Explanation. Retrieved from https://www.thoughtco.com/bell-curve-normal-distribution-defined-2312350
[2] Normal Distribution. (n.d.). Retrieved from https://www.mathsisfun.com/data/standard-normal-distribution.html

The ***mean*** is the average of a set of numbers. To find the ***mean*** of a data set, add up all of the numbers in the set, and then divide that total by the number of numbers in the set.[3]

Standard deviation is a number used to tell how measurements for a group are spread out from the average (mean), or expected value. A ***low standard deviation*** means that most of the numbers are close to the average[4]

2.2 Standard normal distribution

Standard normal distribution is a type of normal distribution where the mean is 0 and the standard deviation is 1. Since a normal distribution is defined by its mean and standard deviation, it can give rise to many bell curves of many different shapes. We can transform each data value into z-score, this is defined as standardisation, hence the data set is fitted into a standard normal distribution, as listed below:

$$Z = \frac{x - \mu}{\sigma} \sim N(0,1)$$

2.3 Z-scores

Deducing from the formula, the z-score also called standard score, tells us how many standard deviations the data value is away from the mean. Z-scores can be positive or negative, it tells us whether our data is above or below average respectively. Its

[3] How to Find the Mean. (n.d.). Retrieved from https://www.mathsisfun.com/mean.html
[4] (n.d.). Retrieved from https://www.mathsisfun.com/data/standard-normal-distribution.html

absolute values tells us exactly how far it is from the mean. This also allows us to compare observations from different data sets.

3.0 Methodology:

In order to carry out my investigation, I had managed to retrieve data with the aid of a high school teacher whom I have known back in China from my hometown. Of these sets of scores, there is one particular student who had scored 80 marks on all subjects, her score is highlighted in each table.

With the agenda to compare exactly how well she did in each subject, I have went ahead and calculated her percentile for each subject. However, she happens to be the 90th percentile for each subject, hence it was inconclusive and a more sensitive statistical measure was needed.

Looking at the Business class of 150 students, with the test scores listed below:

Business Class /Scores														
28	44	50	54	57	60	63	65	67	68	71	73	75	77	85
31	45	50	54	57	61	63	65	67	69	71	73	75	78	87
37	45	52	54	58	61	63	65	67	69	72	73	75	79	87
37	46	52	54	59	61	64	66	67	69	72	73	76	80	88
38	46	53	55	59	61	64	66	67	69	72	74	76	80	88
40	47	53	55	59	62	64	66	68	70	72	74	76	81	88
41	48	53	56	60	62	64	66	68	70	72	75	76	83	90
43	48	53	56	60	63	64	66	68	70	72	75	77	83	91
43	48	54	56	60	63	64	66	68	71	72	75	77	85	92
44	48	54	56	60	63	65	67	68	71	73	75	77	85	93

Table 1: test scores for Business class

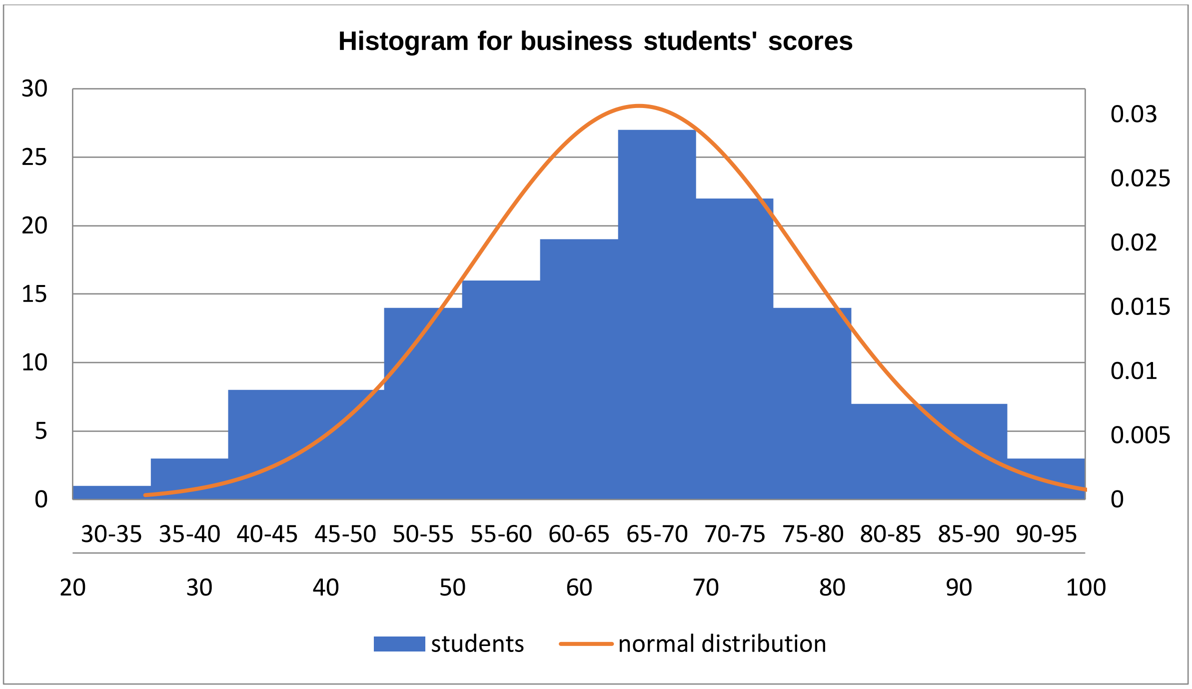


Figure 1: test scores of business students

Calculating mean and standard deviation from the first set of data,

Mean,μ = 65

Standard deviation, σ = 12.965

Let X be the random variable for test scores obtained, therefore,

$$X \sim N\,(65, 12.965^2)$$

After standardisation,

$$Z = \frac{x - 65}{12.965} \sim N(0,1)$$

For the student whom scored of 80 marks, her Z_b would be calculated as follow

$$Z_b = \frac{80 - 65}{12.965} = 1.1570$$

Now, let’s move to the last batch of 139 students, with their test scores listed below:

Economics students’ score													
37	52	57	58	60	62	64	66	67	69	71	73	75	80
39	52	57	59	61	62	64	66	67	69	71	73	75	81
44	53	57	59	61	62	65	66	67	69	71	73	76	82
46	53	57	59	61	63	65	66	68	69	71	74	76	83
48	54	57	59	61	63	65	66	68	69	71	74	80	84
48	55	57	59	61	63	65	66	68	69	72	74	80	84
49	55	58	60	61	63	65	66	68	69	72	75	80	86
50	55	58	60	61	64	65	66	68	70	72	75	80	86
51	56	58	60	61	64	65	66	68	70	72	75	80	89
52	57	58	60	62	64	65	67	69	71	72	75	80	

Table 2: test scores for Economics class

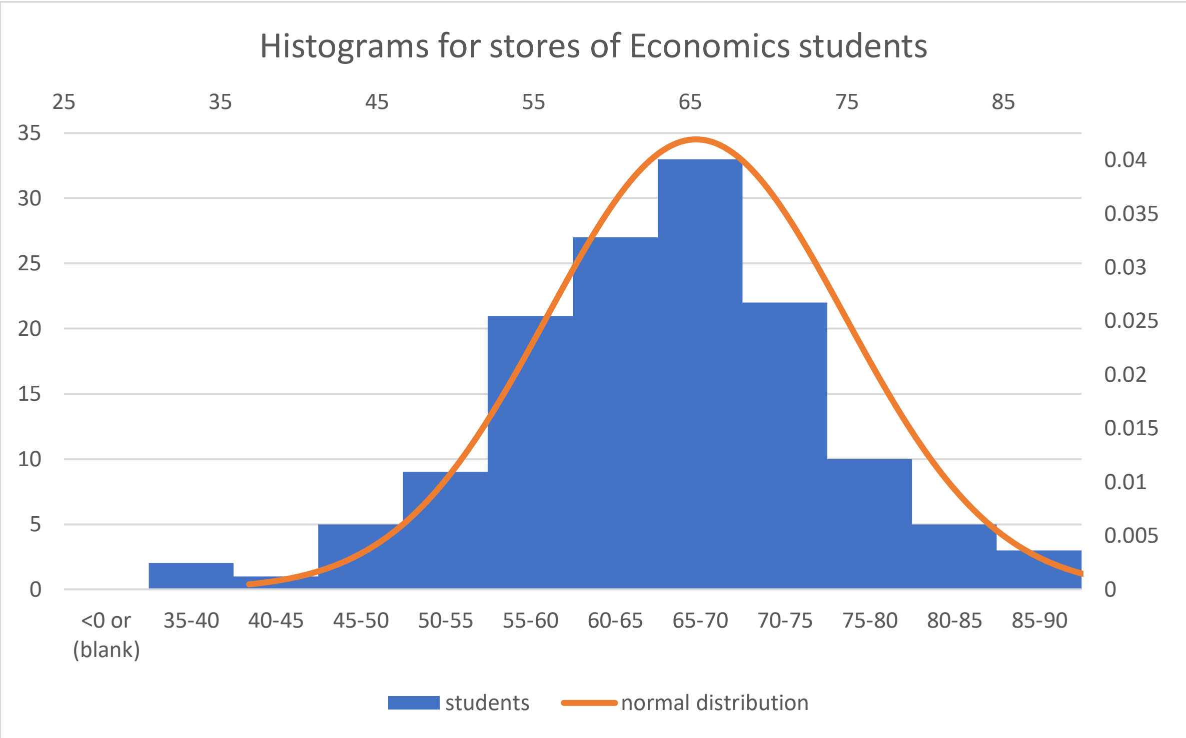


Figure 2: histogram for scores of Economics students

Calculating mean and standard deviation for the from data,

Mean, μ = 65

Standard deviation, σ = 9.4839

Let X be the random variable for test scores obtained, therefore,

$$X \sim N\ (65, 9.4839^2),$$

After standardizing,

$$Z = \frac{x - 65}{9.4839} \sim N(0,1)$$

Z-score for the student who scored a test score of 80 marks on this test, Z_e , would be,

$$Z_e = \frac{80 - 65}{9.4839} = 1.5816$$

Comparing z-scores from the two tests, since $Z_e > Z_g$, it tells us that her score of 80 in Economics test is further away from the average of the class. Thus, we can conclude, with the greater z-score, this student has actually performed better in Economics than Geography as well.

Taking a different approach to analyse the data sets, looking at the mean score of 65 for both sets of data, we can assume that the students had performed similarly well for both tests. However, standard deviation for Economics is 9.4839, which is smaller than that for Business, which is 12.965. This meant that scores for Economics were more consistent, and less spread out from average. Therefore, getting a score of 80 was less likely, and thus we can infer that the student did better for Economics.

<u>Predicting results when mean is different, and standard deviation is the same</u>
Although from the data sets we have, we are not able to investigate what would happen when mean score is the different while standard deviation is the same. From what we know about z-scores, we could predict what we would find it that was the case.

Since z-score tells us the number of standard deviation our data value is from the mean. When standard deviation is the same, the same test score from a class with lower mean would give us a higher z-score value.

For the above two examples, z-score comparison has been done with the assumption that scores from the two tests follows normal distribution. That is, a normal distribution well-models the two data, as shown from the histogram plots.

Next, taking a look at the Geography class of 125 students, with their test scores listed as follows:

Geography Class /Scores												
47	52	55	58	59	63	64	65	68	70	74	80	82
47	52	55	58	60	63	64	66	68	71	75	80	87
47	52	55	58	60	63	64	66	68	71	75	80	87
47	53	55	58	60	63	64	66	68	71	75	80	95
49	53	56	58	61	63	64	66	68	72	75	80	100
49	53	56	59	61	63	65	66	68	72	76	80	
50	53	56	59	62	63	65	66	68	73	77	80	
51	54	57	59	62	63	65	66	69	73	80	80	
51	54	58	59	62	64	65	67	69	73	80	81	
52	54	58	59	63	64	65	67	70	74	80	81	

<u>Table 3: test scores for geography students</u>

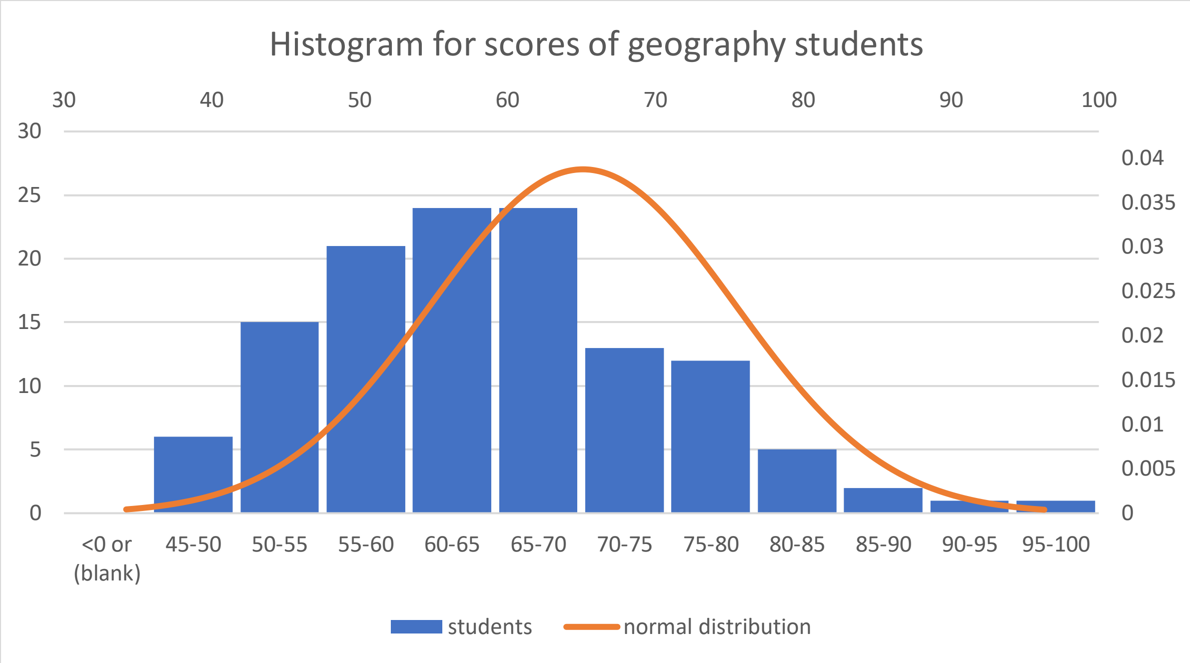


Figure 3: Histogram for scores of geography students

Calculating mean and standard deviation for the from data,

Mean,μ = 65

Standard deviation, σ = 10.258

Let X be the random variable for test scores obtained, therefore,

$$X \sim N\ (65, 10.258^2),$$

After standardizing,

$$Z = \frac{x - 65}{10.258} \sim N(0,1)$$

Z-score for the student who scored a test score of 80 marks on this test, Z_g , would be,

$$Z_g = \frac{80 - 65}{10.258} = 1.4623$$

Comparing z-scores from the two tests, since $Z_e > Z_b$, it tells us that her score of 80 in Economics test is further away from the average of the class. Thus we can infer that this student has actually performed better in Economics than Business.

Making comparison with skewed distribution

Before concluding that the student has performed best for Economics compared with Geography and Business. I have noticed that distribution for Geography test scores is moderately right skewed. As mentioned, comparison of z-score for Business and Economics test were done based on assumption that normal distribution well-models the score distributions respectively. From what I know, for a skewed distribution, both the mean and the standard deviation are affected by the skew in such a way that makes the z-score results less indicative of what it is trying to convey.

In fact, considering the scenario **if** mean scores and standard deviation for Geography is equal to those of Economics, we can predict that z-score for test score of 80 would be the same. That is,

$$Z_e = Z_g = \frac{80 - 65}{9.4839} = 1.5816$$

However, since distribution is skewed, I feel that we cannot safely conclude that the student had performed equally well for both tests based on z-score.

So the question arises, when percentiles are the same, mean and standard deviation, and hence z-score are equal, how should we compare performance if the distribution of the data is skewed.

To deal with skewed data, statistician would usually normalise their collected data to ensure it fits well into a normal distribution. The idea of normalisation is to put all raw data through a function, such that the transformed results would follow a normal distribution more closely. This is done because many statistical tests and results are performed with the assumption that data follows a normal distribution.

A right skewed distribution can be normalised using *logarithm transformation* which can effectively reduce the skewness of the data.

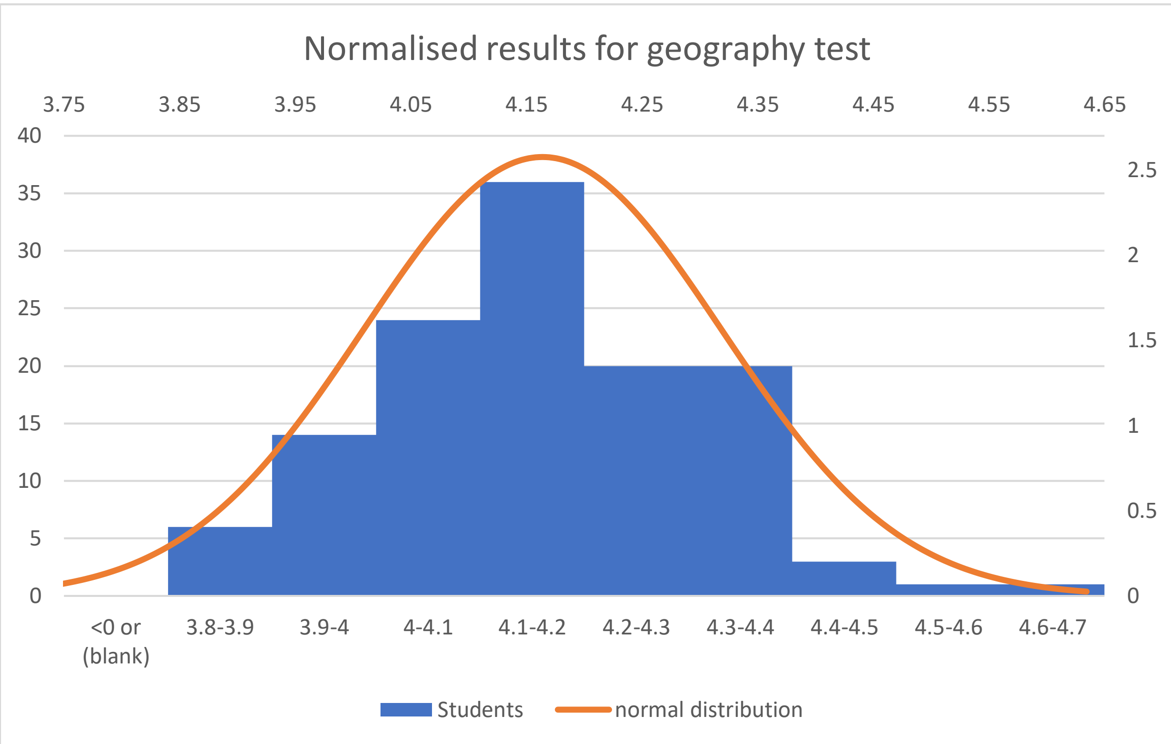


Figure 4: Histogram for normalised results

Taking natural logarithm on all data values,

$$\ln X \sim N(4.16, 0.15489^2)$$

After Standardising,

$$Z = \frac{\ln x - 4.16}{0.15489} \sim N(0,1)$$

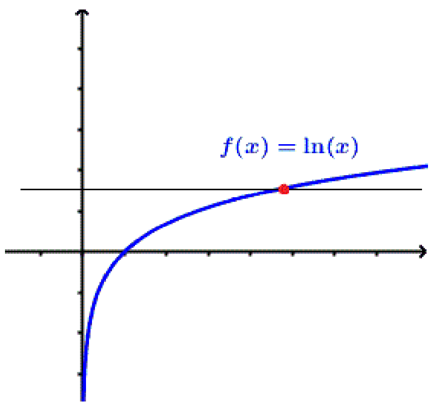
Calculating the z-score for transformed value, $Z_{g(t)}$

$$Z_{g(t)} = \frac{\ln 80 - 4.16}{0.15489} = 1.4334$$

When comparing z-score of transformed data after using logarithm transformation, the resulting z-score reduces. Logarithm function affects larger values more, in the sense that larger values reduces more significantly than low values. When the we take natural logarithm on test scores, the higher score will be reduced more, in contrast, to the lower score that will be less affected. This benefits the poorer performing students giving them better grades.

As shown in the logarithm equation:

$$Y = \ln x$$



From the graph, as x tends to positive infinity y tends to positive infinity but plateaus off. When a set of skewed data is transformed to better suit normal distribution, care should be taken, considering the underlying parameter of the data.

4.0 Conclusion

In conclusion, z-score comparison allows us to make comparisons between different distributions. When the distribution is normal, standardizing to z-score places each data value on a standard normal distribution which has a mean and standard deviation of 0 and 1 respectively. Hence this allows us to compare each z-score on the same scale, even if they came from different distribution. That is, comparing an apple with an orange.

In this case, it may not be easy to state which subject a student has performed better in simply looking at the raw score from the tests. By looking at where they stand in each of their distribution, we can meaningfully infer in which test the student performed best in.

Approaching to normalising data to better fit a normal distribution model should also be done with care. Although is it a common procedure for many researchers, we note that in this case, applying to test score may not be suitable. This is due to the nature of what test scores measures, and how it is affected by a logarithmic function. In cases where distribution of scores are skewed, care should be taken to make a logical comparison than simply following normalization procedures.

5.0 Evaluation

The limitations of the investigation are that distribution of test scores may not exactly be well modelled by the normal distribution. Goodness of fit of the distribution model comes into question. In reality, when test scores do not model a bell curve, teacher may opt to use another form of grading such as percentage grading. It is because one assumption was made that all the students' scores follow normal distribution.

From what I know, when distribution is skewed. Mean may not accurately describe the central tendency of a data. For the Geography test, a right skew distribution of scores implies to me that the test could have been hard, that’s why more students scored a lower mark and few score high marks. In this case, modal class of 60-65 should reflect average of the class better. Also, as mentioned, since high marks are scarce, scoring 80 suggests to me that the student had performed pretty well. Hence, I feel it would have been more logical to conclude that the student had actually performed better for Geography than Economics.

6.0 Real world applications

Z-scores can be used to examine the variability of a data and it is applied by dealers to predict any changes in market. It is commonly known as the Altman z-score when it is used as a credit-strength test that measures the probability of bankruptcy for a public trading company. Z-score can be calculated using the five key financial ratios: Working capital, Retained profits, Profit before interest and taxes, market value of equity and sales which can be found in the company’s annual report.

$$\boldsymbol{\zeta = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E}$$

$$\zeta = the\ Altman\ score$$

$$A = Working\ capital$$

$$B = Retained\ profits$$

$$C = profits\ before\ interests\ and\ tax$$

$$D = market\ value\ of\ equity$$

$$E = Sales\ revenue$$

Similar to student scores, positive or negative scores indicates whether the company is in the “safe” or “distress” zone. The absolute value positions where the company are. Any score below 1.8 specifies that the firm is most likely experiencing insolvency. On the other hand, if the firm has a score that is above 3, it indicates that they are having a good cash flow and are not likely to be bankrupted.