**Draft writeup for Matt:**

A critical review of the 2013 group’s 2013 group’s statistical frequency analysis was conducted and can be seen in Method – Specific Tasks, Task 1 section.

Background research in Java for use in compiling and editing 2013 group’s statistical frequency analysis code

Preparation of test text files from the Universal Declaration of Human Rights, which were then processed in Java and the letter frequency results were used to test the P-value code in Excel, Matlab and by hand.

Firstly, a group of test cases were prepared, using 3 groups of 44 letters from the English, French and German languages from the Universal Declaration of Human Rights. These were analysed using ANOVA in Excel,

Once these results were obtained, the P-value calculation method used was tested using various software. A test case with known p-value result from the Engineering Maths IIA notes was run through ANOVA in Excel, using built in Matlab Functions found in Engineering Maths IIA notes, and by hand.

Testing of P-value method using various software etc.

I tested using Excel and Matlab

Jikai tested using Excel and by hand

Tried using ANOVA One Way comparing Letter Frequency between Base text and code, and ANOVA Two Way Analysis of Variance, where Factor 1 was the Letter, Factor 2 was the Text and the Response was the Letter Frequency. Both did not work, ie. Uses comparison of total means to determine p-value

Researched and tried paired sample t-test, ---- Tried using paired data analysis from Engineering Maths IIA Notes and Matlab, did not work, ie. Uses mean and standard deviation of the differences between samples to calculate the p-value, thus letter and text type data is lost in the process and so is not applicable.

All of these methods do not compare individual data entries from one group and their corresponding entry in another group. Instead they use the sample size, sum, mean and variance of each group to compare to the other group. Because of this, this method cannot be used when comparing letter frequency between languages. For example the frequency of the letter A in one text, must be compared with the frequency of the letter A in another text, rather than the frequency of all letters in one text being compared to the frequency of all letters in another text.

Also tried linear regression and correlation methods of calculating p-value but neither of these worked either.

Researched and tried paired sample t-test, still not applicable since uses sample means for comparison.

The rest of the text files from the Universal Declaration of Human Rights was obtained from the 2013 group, processed using a modified version of the 2013 groups initial letter frequency code, and the results tabulated in an excel spreadsheet including language, numerical frequency of each letter in each language, percentage frequency based on the total number of initial letters in each language, and numerical and percentage frequency of letters in the Somerton Man code. This data spreadsheet is to be used for statistical calculations for comparison of the letter frequency in the code versus each language in the Universal Declaration of Human Rights.

Questions:

It seems as though a p-value analysis cannot be done to compare texts as for example if you have a text with 5 As and 0 Bs, and a second text with 0As and 5Bs, they will have the same mean of 2.5 and thus produce a perfect p-value of 1, meaning an exact match, when in actuality they are the complete opposite. This happens regardless of whether using a paired method or regular method.

Do you know of any way we can still obtain a p-value using any other method? I have tried linear correlation and regression but these give mixed results, in the examples in the Linear Correlation and Regression excel spreadsheet, you can see that Zapoteco achieves a much higher p-value in both cases than English, when by observation it looks to be far more different to the code than English.

Otherwise, do you know of any other ways of comparing the letter frequency between texts? I have attached the 2013 group’s method of using average standard deviation and average squared difference for reference.

Thanks for your help Matthew