**Department of Electrical and Computer Engineering**

Homework Assignment No. 07:

**HW No. 07: Information Theory and Statistical Significance**

submitted to:

Professor Joseph Picone

ECE 8527: Introduction to Pattern Recognition and Machine Learning

Temple University

College of Engineering

1947 North 12th Street

Philadelphia, Pennsylvania 19122

March 8th, 2022

prepared by:

Gavin Koma  
Email: [gavintkoma@temple.edu](mailto:gavintkoma@temple.edu)

# Task 1

The goal of Task1 is to assess a two-dimensional dataset with feature vectors: [*x­­*1, *x*2]; and then spend time to quantize each element of each vector to a set of 128 discrete values. We are given the equation do this already and is as such:

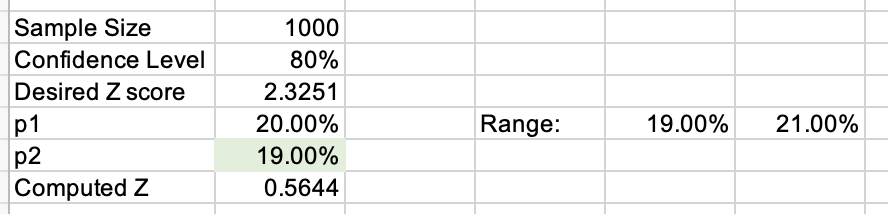
In order to do so, I first began by::: ??????? idk bc I haven’t done this yet lol

# Task 2

Task 2 involves assessing two systems (a baseline system and a new system) which assess performance of a dataset of 1,000 files. We are initially told that the baseline system gives us an error rate of 20.0% and the new system delivers and error rate of 19.0%. Our first goal is to determine if the new system, which delivers an error rate of 19.0%, is statistically significant at a confidence level of 80%.

We can do this by assuming that P1 and P2 are 20.0% and 19.0%, respectively. It is also important to note that the value of N will be the same: 1,000 total datafiles. From here, we can utilize the Z-score formula which is as follows:

We are also provided with an Excel sheet that is capable of calculating our Z-score given a value for P1 and P2. We are able to see that the calculated computed Z score from this Excel sheet is the same as the Z score calculated from the above equation. Attached below is a screenshot of the Z-score.



As discussed in class, we know that any positive scores that are in the z-table will correspond to some value which is less than the mean while the negative scores in the z-table will correspond to any values that are less than the mean. Recall that our calculated z-score is 0.5644 and we can utilize the following formula:

Most important to recall is that the above calculation functions for a one-tailed t-test. Since we do want to consider the effects and consequences of missing an effect in the other direction and thus, we should multiply the area by two.

This calculation allows us to find the z-score according to a corresponding area. In this case, the z-score for area (0.2) is less than the computed z-score found above (0.5644). In this case, we know that the z-score which has been calculated by the first formula is larger than the value that is calculated by the second formula. As a byproduct, we are able to say that the new system error rate of P2 is notably smaller than that of P1 but is still not statistically significant as it is larger than 0.05.

The next portion of this assignment asks us to assess what the minimum decrease in error rate that will result in a statistically significant result. To make things easy, we can assess the various decreases in error rate that may result in something statistically significant.

Reconsider the following hypothesis:

We first need to pick to assess varying error rates for the base system and the new system. This means that we need to assess when the difference will result in a p-value that is less than 0.05. Our critical value is our calculated z-score. We can reject the null hypothesis when the critical value is less than our p-value or when c < mean. At an error rate of 0.1992, our critical value is 0.0448 which is less than our p-value and thus we can reject the null and note statistical significance in our data.

Our final goal of this task is to repeat the first two tasks for *N* = 100, 500, 2000, 5000, and 10000 with a confidence level of 85%, 90%, and 95%.

# Conclusion