**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:

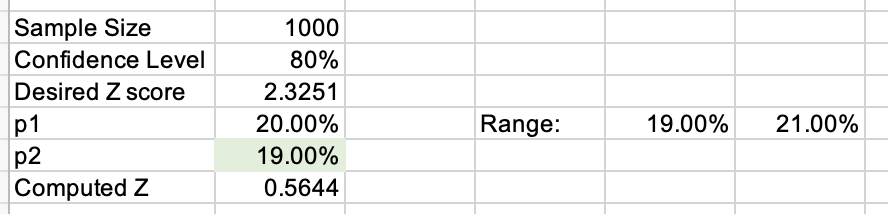
In order to do so, I first began by::: ??????? idk bc I haven’t done this yet lol

# 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:

This calculation allows us to find the z-score according to a corresponding area. In this case, the z-score for area (0.1) and the corresponding desired score for this area is 1.2820 which is larger than the computed z-score found above (0.5644). When we compare these two z-scores, we can make the decision for rejecting our null hypothesis. Let us state the following null and alternative hypothesis:

For part 1 of this Task, we can compare the two z-scores (1.2820 and 0.5644) and we can say that the z-score calculated via formula is larger than the z-score calculated by the area formula. We can state that P2 (the new system error rate) is significantly lower than P1.

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.

To do this, we can utilize the Excel calculator provided and input varying numbers until we see a result that is *not* statistically significant. We are able to determine that a decrease in error rate of 00.17% will result in a not statistically significant result. Such a decrease will result in a z-score of 0.0952 which is less than our area z-score calculation and we are able to determine that P2 (the new system error) is not statistically significant when compared to P1 (the old system error)

Table

Description automatically generated

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 80%, 85%, 90%, and 95%. Let us first refresh our memories as to what our chosen hypotheses are:

For ease of use, I wrote a simple python function that will calculate the score of both area z and computational z. This is just so I don’t have to continue clicking through Excel but instead can just alter the variables. The equations are the same as utilized in the Excel sheet, so I have elected to not include a screenshot of the code.

Again, we will want to compare the area score to the other score. The scores are compiled below.

Table

Description automatically generated

Table

Description automatically generated

We see that the minimum rate of decreases as we increase the confidence interval. Given lower confidence intervals, we are giving room for more error in our system while a smaller confidence interval decreases the room for error. We see this trend through all of the included charts above and it is an important thing to note for this task. We are able to determine significance by comparing the computed z-score to the area z-score that can be found in a z-table. We are able to reject our null hypothesis that there is no difference between the two systems, and we are able to state that there is some difference given a certain error rate. To do so, if our computed z-score is *less* than the desired z-score (found in the table) then we can reject our null hypothesis.

# Conclusion