Robert Irwin

ECE 3522: Stochastics

Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA 1912

# Problem Statement

In this assignment we will be doing hypothesis testing. Hypothesis testing is a tool used to test a claim based off of a given mean and a given standard deviation. Whether or not the claim is valid depends on the confidence level you are testing to. As the confidence level increases, the range of acceptance increases, because you are more certain of the claim.

We will be using the following formula to calculate the z scores, which are compared to the critical z-scores that correspond to certain confidence levels. It is shown below in Equation 1.



In this assignment we will be generating a set of Gaussian distributed numbers with a specified mean and standard deviation, and testing a companies claim based off of the mean of the set, and sample sizes of n = 10,100, and 1000. If the difference in the means is not statistically significant, we will calculate the minimum sample size, n, that would make the difference of the mean statistically significant.

# Approach and Results

The first step in the analysis was to create the Gaussian distributed sample set. This will be discussed further in the description of the Matlab code. The results for difference confidence levels, and different sample sizes are shown below. Significance was determined using the z-score analysis shown in Equation 1.



*Figure 1: Significance for n=10 stdev = 120*

Each column represents a confidence level, from left to right, starting at 80% and ending at 99.99% confidence. A one means that the difference in the mean was not statistically significant. For this population size, we see that the difference in the means were not significant. This is not surprising, as we generated the 10 samples to have a mean and standard deviation equal to the claim of the company, i.e. mean = 1600, and stdev = 120. Now we will look at the minimum population size to make the difference in the mean significant.



*Figure 2: Minimum n value for significance*

Again, the columns refer to different confidence levels. The values make sense. Notice that n is very large compared to the sample size of n. In order for the difference in the means to be statistically significant, the population size that has a mean different from the expected mean needs to be large. It also makes sense that the minimum n increases for increasing confidence. This is because if we are extremely confident in our claim, we need a lot of data that tells us our claim is invalid before we believe it.

For a sample size of n = 10, the mean of the sample population is seen below.



*Figure 3: Mean of Sample Population*

Notice that the mean is very close to the desired mean of 1600. This is why it makes sense that at no confidence level was the difference in the means statistically significant.

The same results will be shown for a sample size of n = 100.



*Figure 4: Significance for n=100 stdev = 120*

Note that a zero means that the difference in the means was statistically significant. The columns refer to the same thing as described above.



*Figure 5: Minimum n for Significance*

For the first two columns, a minimum n was not calculated because the difference in the mean was statistically significant. When the difference in the means was not statistically significant, our calculated n was higher than the population size. This makes sense for the same reasons described above. The actual mean of the set is presented below.



*Figure 6: Actual Mean for n = 100*

Here we see that the actual mean is pretty far from the expected mean. For this reason, we expect some statistically significant data, which we observed.

Now we present the same observations for n = 1000.



*Figure 7: Significance for n = 1000 stdev = 120.*

We see that at every confidence level tested, there was no statistical significance in the difference of the means.



*Figure 8: Minimum n value for Significance*

Again we see that the value of n is larger than the population size, and the minimum n increases at higher levels of confidence.



*Figure 9: Actual Mean for n = 1000*

Our results make sense, as the actual mean is very close to the expected mean.

We now repeat the process for a standard deviation of 240. Because the analysis is identical, all results will be presented, and discussed at the end.



*Figure 10: Significance for n=10 stdev = 240*



*Figure 11: Minimum n for Significance*

**

*Figure 12: Actual Mean for n = 10*

**

*Figure 13: Significance for n=100 stdev = 240*

**

*Figure 14: Minimum n value for Significance*

**

*Figure 15: Actual Mean for n = 100*

**

*Figure 16: Significance for n=1000 stdev = 240*

**

*Figure 17: Minimum n value for Significance*

**

*Figure 18: Actual Mean for n = 1000*

An important thing to notice here is that the actual mean of the sample population can be much further from the expected mean. This is because we expect more variation in the products. This larger variance in the products is a direct result of increasing the standard deviation.

For the final part of the experiment, when we were trying to find the maximum variance where the difference in the means was significant, we got no conclusive results. Even at a standard deviation of 1 trillion, we still got some statistically significant results. This will be discussed further in the conclusion section of the report.

# MATLAB Code



*Figure 19: Defining Constants and Plotting*

Here we define our confidence levels, expected mean, standard deviation, and sample sizes. We then pass these parameters into the user defined function gen\_grv(), which will be described below.



*Figure 20: Gen\_grv()*

This function takes the mean, standard deviation, and sample size and generates a Gaussian distributed set of data using the box-Muller approximation.



*Figure 21: Checking and Determining the Significance*

In this code we calculate the mean of each sample size, and pass it to check\_significance(). We pass the same parameters to determine\_significance(). Both are user-defined functions, and will be explained below.



*Figure 22: Check\_significance()*

This function takes the actual mean, expected mean, standard deviation, sample size, and confidence level. Based off the confidence level, we identify the critical z scores. Then we calculate our z score using Equation 1.



*Figure 23: Check\_significance(continued)*

Once we have our z score, we check to see if it is in the acceptance range. If it is, we set status equal to 1, which says the difference in the means is not statistically significant. If this is not true, status is set to 0 which means the difference in the means is statistically significant.



 *Figure 24: Determine\_significance()*

This functions takes the same parameters as check\_significance(). It assigns the critical z scores in the same way as check\_significance as well. It then tests to see if the difference in the means is significantly significant. If it is not, it calculates the minimum n required for statistical significance. We do this by manipulating Equation 1 to get n by itself. We then set z equal to the critical z-score to find the n that puts our mean at the edge of the acceptance region.



*Figure 25: Code for Part 5*

The code for part is the same as described above, except it is looped for every specified standard deviation.

# Conclusions

In this assignment we have made many important observations about hypothesis testing. The first is that the higher your confidence level, the wider your acceptance region becomes. We have also seen that the acceptance region expands with increasing standard deviation. The population size also impacts the significance of the difference of the means. For larger n’s the data has more of a chance of the difference in the means to be statistically significant. The reason why we could not find a maximum variance where the difference in the means was statistically significant is because we are generating a Gaussian distributed set based off of the specified standard deviation. For this reason, there is always a possibility that the difference in the means will be statistically significant. If there is a maximum variance that achieves this, it is unrealistically large. We tested standard deviations up to 1 trillion and still found statistically significant results for all sample sizes. All of these calculations were based off a random number generator, so the results are somewhat unpredictable, unless you can generate infinite samples!