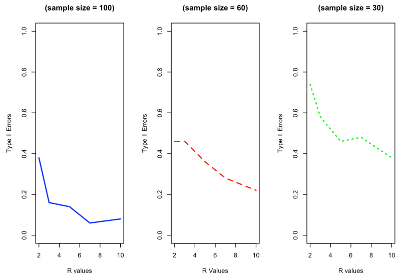
stat428report

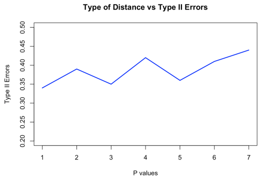
2023-12-10

Introduction: A pharmaceutical company would like to test whether the effects of two treatments are similar or not. The manager wants to choose two sample testing methods among the four. There are advantages and disadvantages to each test and the purpose of this report will be to illustrate those advantages and disadvantages and provide a recommendation on the test to utilize as well as provide the reasoning behind doing so. The nearest neighbor test uses neighboring data points to examine the closeness of the observed points compared to expectation. The energy distance and Hotelling's T square Test uses distance metrics to quantify the difference between the two samples and the graph based test compares the two samples by creating a graph.

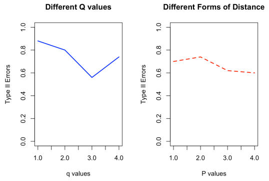


Nearest Neigbor Parameter Selection

We will be using the type II error when evaluating these tests. In layman's terms, type II errors refers to false negatives or the probability of concluding that the two samples are different when in fact they are not. The first decision when implementing these tests would be selecting the optimal choice of parameters or settings of the test. For Nearest Neighbor Test as seen from the plot below, in general as the number of neighbors increases, the type II error decreases and this also varies by the sample size as shown below. As a result, the desired type II error and number of data points influences the number of nearest neighbors needed.



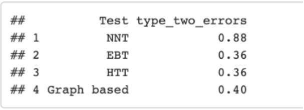
Energy Distance Test Parameter



Graph Based Test Paramaters

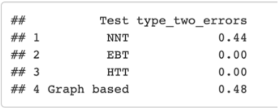
For the Energy distance test, it is slightly harder to find an optimal distance form which is one of the parameters. As seen from the plot on the right for Energy distance test, there is not a clear trend on what the optimal distance would be. This might be one of the disadvantages of the energy distance is that it requires more effort in finding the optimal parameter. Similarly, for graph based tests as seen above, finding optimal parameters for q and distance form may require more effort as the trend between Type II and the parameters is not as clear as that from Nearest Neighbor Test. However, overall type II error does influence both the parameters and can be used to select parameters in the test for both Energy distance and graph based tests.

All in all, when using Nearest Neighbor, it may be easier to select the optimal parameters due to a clear trend between type II error and the nearest neighbors while it may be harder to select optimal parameters for the other tests mentioned above.



```
##          Test type_two_errors
## 1         NNT            0.88
## 2         EBT            0.36
## 3         HTT            0.36
## 4 Graph based            0.40
```

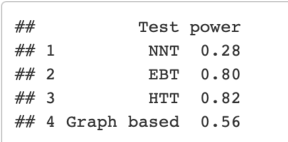
Low Dimensional Results



```
##          Test type_two_errors
## 1         NNT            0.44
## 2         EBT            0.00
## 3         HTT            0.00
## 4 Graph based            0.48
```

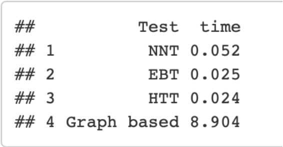
High Dimensional Results

For the graph based test at lower dimensions (dimension = 2), the type II error is 0.40 and at higher dimensions (d = 7), the type II error increases to 0.48. For graph based test, type II error is less at lower dimensions than higher dimensions. This means that graph based test is more suitable for low dimensional data. As seen from the tables above, NNT, EBT, and HTT all have lower type II errors at higher dimensional data than lower dimensional data, suggesting that those tests could be more suitable for higher dimensional data.



```
##          Test power
## 1         NNT  0.28
## 2         EBT  0.80
## 3         HTT  0.82
## 4 Graph based  0.56
```

Power of Test



```
##          Test  time
## 1         NNT 0.052
## 2         EBT 0.025
## 3         HTT 0.024
## 4 Graph based 8.904
```

Computational Efficiency of Tests

The power of each test is shown above. From our simulation results above, HTT is the most powerful test out of all 4 tests. EBT is also powerful while NNT is the least powerful test. In layman's terms, the power of a test refers to the likelihood of detecting a difference between the samples if there is one. In other words, from the experiment results, HTT has the highest probability of detecting a difference between the samples if there is one. From the simulation results above, HTT is the most computationally efficient test because it took the least amount of time to execute the code for the test. Next in terms of efficiency would be NNT and EBT. Finally, graph based test is the least efficient as it took by far the most amount of time for the test to execute. In this case, efficiency refers to the length of time it would take to run the test.

In general, the recommendation for the type of test to utilize would be HTT. First, HTT is the most computationally efficient out of all the tests. This will make it easier for the pharmaceutical to implement and put into practice in a production environment. It will take fewer resources, time, and energy to conduct this test and it will be more easily scalable to larger datasets due to its computational efficiency. Additionally, one of the advantages of HTT is that it does not require selecting optimal parameters as we have seen above for the other tests. This can save time, energy and effort in running and implementing the test. HTT is the most powerful test as seen from the simulation results. This means that the test has a greater likelihood of detecting a difference between the samples if there is one. For example, HTT was able to detect a difference between the two samples in the ElemStatLearn/datasets/prostate.data in the overall difference between people older than 65 (age>65) and younger than 65 (age<=65).