Computer Vision (CS 419)
Project Report on

"Handling Data Scarcity in Training Deep Neural Networks through Data Augmentation"

By:-
Arushi Jain (160001008)
Priyanka Rotte (160001051)

Under the Guidance of
Dr. Surya Prakash
Department of Computer Science and Engineering
Indian Institute of Technology Indore
Spring 2020

Introduction

Nowadays, most of the object recognition techniques are using 3D data instead of 2D. The performance on 3D data is significantly better than that on the 2D data. For example, in the case of face recognition, 2D face recognition is hindered by pose, expression, and illumination variations. These limitations are overcome when using 3D data as all the information about the face geometry is processed in the 3D based approach.

Although the 3D object recognition achieves great accuracy, 3D data acquisition from objects takes time and hence often, there is very limited data available for 3D objects. In the availability of limited data, the model learns the details and noise of these few samples so well that it negatively impacts the testing of the selected model on new data. To avoid this problem of overfitting, we need to increase the variability of the 3D data by enlarging the size of the database by making use of data augmentation.

3D input data for an object, which is in the form of a point cloud, contains an unordered set of 3D points. It is seen that this original set of 3D points for an object contains a huge number of 3D points; however, due to the computational and memory limitations of the system, often, we cannot use the entire point cloud of a single sample for processing. To mitigate this problem, usually, the original point cloud data is sub-sampled, and a reduced size cloud is used for processing. However, in this process, the number of samples for a subject remains the same as was available earlier before sampling. We can exploit the use of sampling in a different way and propose its use in data augmentation by increasing the number of samples of the subjects.

In this project, we propose three sampling techniques that can be used for creating subsamples from an original sample. We use the ICP (Iterative Closest Point) algorithm to show that the samples created from the original data all carry the same information. Then, we use Central Limit Theorem to prove that the information carried by the subsamples is the same as that carried by the original sample i.e. they have the same discriminative power. Finally, we compare the three sampling techniques based on the results.

Methodology

Given a 3D point cloud, we use the following types of sampling techniques for data augmentation:

- ➢ Random Sampling:
  In random sampling, to create multiple subsamples from a single sample, we randomly select a fixed proportion of points from the original point cloud multiple times. This creates different unordered subsets containing a uniform number of points. In our experiment, we are selecting one-third of the original points for each sample.

- ➢ Systematic Sampling:

In this technique, we sort the point cloud of a sample by ordering the points in 6 possible arrangements - (x, y, z), (x, z, y), (y, x, z), (y, z, x), (z, x, y), (z, y, x). For each arrangement, we choose a random starting point in [0, k-1] and choose the subsequent points after skipping k, 2k, 3k... points where k lies in [3,5] depending on how crowded or sparse we want our subsamples to be. Lower k results in a lower variance of points among different subsamples while higher k results in less repetition but sparser point clouds. However, we need to ensure that the chosen k isn't symmetric about the point cloud as this will result in the same subsampled point cloud irrespective of the ordering arrangement. In our experiment, we are using k = 3 so that the subsamples use one-third of the point cloud.

➢ Stratified sampling:
In this technique, we divide the entire point cloud of a sample into cubical windows of fixed size and then select a proportionate number of points randomly from each window to create a single subsample. Hence, a higher number of points are selected from a dense region, and a lower number of points are chosen from a sparse region thus maintaining localization. In our experiment, we are using a window size of 5*5*5, and we are selecting one-third of the total points from each window.

We use ICP and Central Limit Theorem to prove that the subsamples created from our original samples all carry the same information and that they have the same discriminative power as the original sample respectively.

★ Iterative Closest Point algorithm: The ICP algorithm finds the transformation matrix between two point clouds by minimizing the square errors between them. One of the point clouds (target) is fixed, and the other one (source) is transformed to best match the target. The algorithm is iterative and improves the transformation matrix to minimize the error. Finally, it returns the final error after the transformation and the transformation matrix. The error is essentially the registration error between the two point clouds.

★ Central Limit Theorem: The central limit theorem says that for any kind of data with a high number of samples:
    - Sampling distribution's mean should be equal to the population mean
    - Sampling distribution's standard deviation should be equal to the population standard deviation divided by the square root of the total number of samples.

Experiments and Results

A. Experiment I

We use 3 samples and for each sample, we create 3 subsamples.
Table 1.1 shows the ICP registration error between a given sample and its respective subsamples for all 3 samples. Table 1.2 shows the ICP registration error between all 3 pairs of subsamples of each of the 3 samples.

Random				
Table 1.1	Subsample 1	Subsample 2	Subsample 3	Mean
Sample 1	4.18E-08	3.49E-08	3.66E-08	3.78E-08
Sample 2	0.00E+00	0.00E+00	8.05E-09	2.68E-09
Sample 3	1.83E-08	2.72E-08	2.17E-08	2.24E-08

Systematic				
Table 1.1	Subsample 1	Subsample 2	Subsample 3	Mean
Sample 1	3.73E-08	3.73E-08	3.32E-08	3.59E-08
Sample 2	1.13E-08	1.80E-08	0.00E+00	9.77E-09
Sample 3	1.83E-08	1.41E-08	1.83E-08	1.69E-08

Stratified				
Table 1.1	Subsample 1	Subsample 2	Subsample 3	Mean
Sample 1	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Sample 2	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Sample 3	0.00E+00	0.00E+00	0.00E+00	0.00E+00

Random				
Table 1.2	Subsample 1&2	Subsample 2&3	Subsample 3&1	Mean
Sample 1	7.89E-01	7.89E-01	7.82E-01	7.86E-01
Sample 2	7.82E-01	7.95E-01	7.85E-01	7.87E-01
Sample 3	7.87E-01	7.92E-01	7.85E-01	7.88E-01

Systematic				
Table 1.2	Subsample 1&2	Subsample 2&3	Subsample 3&1	Mean
Sample 1	7.47E-01	7.35E-01	8.17E-01	7.66E-01
Sample 2	7.50E-01	7.38E-01	8.34E-01	7.74E-01
Sample 3	7.43E-01	7.34E-01	8.28E-01	7.68E-01

Stratified				
Table 1.2	Subsample 1&2	Subsample 2&3	Subsample 3&1	Mean
Sample 1	7.87E-01	7.89E-01	7.85E-01	7.87E-01
Sample 2	7.84E-01	7.84E-01	7.84E-01	7.84E-01
Sample 3	7.82E-01	7.80E-01	7.83E-01	7.82E-01

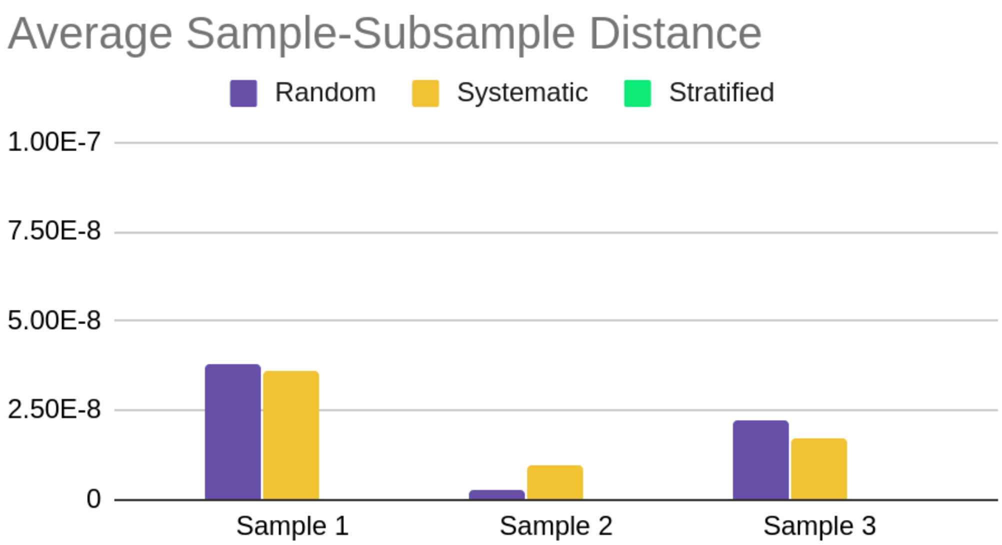

Table 1.1

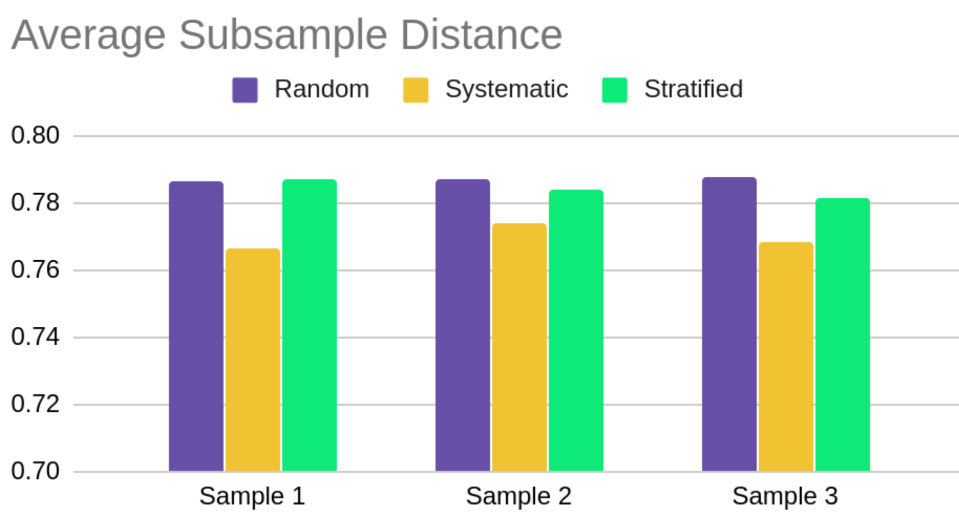

Table 1.2

INFERENCE: From table 1.1, we can see that the registration error between the original sample and its sub-samples is very similar for all the sub-samples proving that all the sub-samples of a particular sample carry the same information.

From the first graph, it is evident that the sample-subsample similarity is maximum in the case of stratified sampling which can be explained because of the use of localization in selecting points. Systematic sampling has the next best similarity owing to ordering of the points before selection. Random sampling has the highest error because of the absence of any ordering or localization.

From the second graph, we can see that the subsample similarity is highest in the systematic sampling because in the systematic sampling technique, there can be repetition as we might choose the same set of points which is not desirable. For an effective sampling technique, we need low subsample similarity for more variation in the data after the augmentation.

B. Experiment II

In this experiment, we are taking 2 different samples of the same subject - Sample 1 aTable 1.2 shows the ICP registration error between all 3 pairs of subsamples of each of the 3 samples.d Sample 2. We also take the respective subsamples namely subsamples 11, 12, and 13 from sample 1 and subsamples 21, 22, and 23 from sample 2. Table 2 shows the ICP registration error between sample 1 along with its respective subsamples and sample 2 along with its respective subsamples.

Random					
Table 2	Sample 2	Subsample 21	Subsample 22	Subsample 23	MSE
Sample 1	8.4114	8.3418	8.4907	8.44	0.05465942279
Subsample 11	8.4472	8.3779	8.5268	8.4753	0.07036380462
Subsample 12	8.4466	8.3765	8.5258	8.4756	0.07011784723
Subsample 13	8.4452	8.3761	8.525	8.4735	0.06919158186
MSE	0.03026185718	0.04590890437	0.1067684059	0.0567433256	

Systematic					
Table 2	Sample 2	Subsample 21	Subsample 22	Subsample 23	MSE
Sample 1	8.4114	8.4109	8.343	8.3984	0.03481310816
Subsample 11	8.4427	8.442	8.3743	8.4299	0.03014427806
Subsample 12	8.4421	8.4416	8.3738	8.4293	0.02995287966
Subsample 13	8.4428	8.4418	8.3739	8.4299	0.03024326371
MSE	0.02696358656	0.02632873905	0.04710355613	0.01713118501	

Stratified					
Table 2	Sample 2	Subsample 21	Subsample 22	Subsample 23	MSE
Sample 1	8.4114	8.4079	8.3954	8.3869	0.01473516203
Subsample 11	8.447	8.4438	8.4308	8.4227	0.02655734362
Subsample 12	8.4453	8.4411	8.42898	8.4212	0.02467993314
Subsample 13	8.4446	8.4409	8.4283	8.4202	0.02416371246
MSE	0.02965977916	0.02652522384	0.0175144683	0.01501182867	

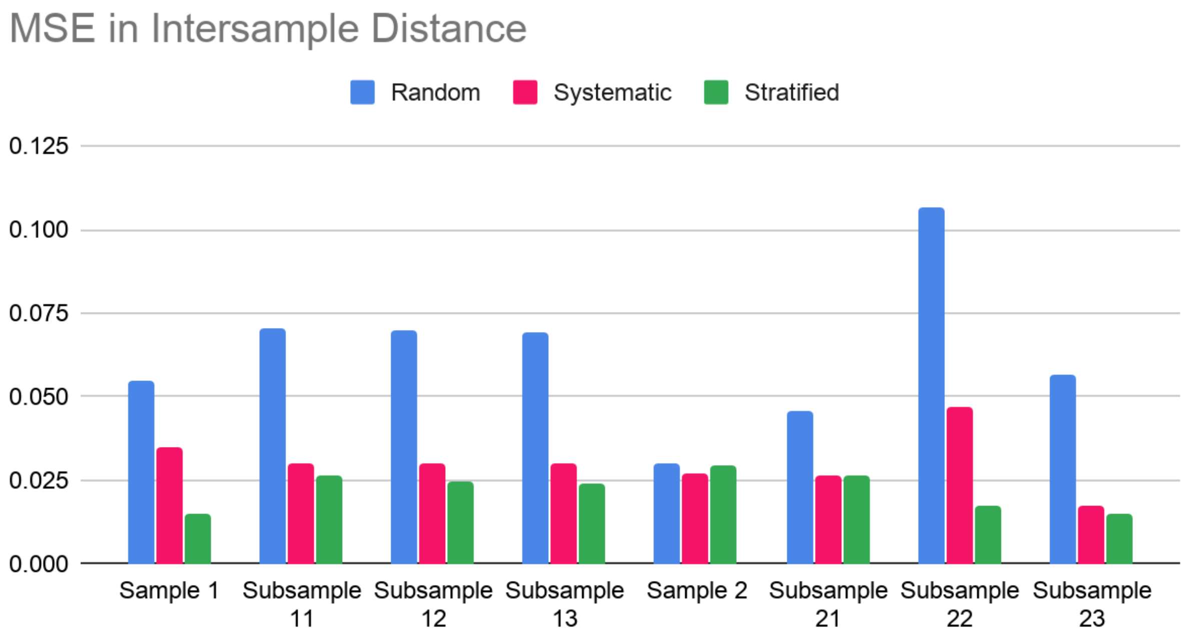

Table 2

INFERENCE: From the above experiment, we can see that the registration error using the subsamples is very similar to that using the original samples. The error is lowest in the case of stratified sampling as this sampling makes use of localization while selecting points. Random sampling has the highest error values among the three as the points are selected randomly from the point cloud without any specific ordering or localization, whereas ordering is taken into consideration in case of the systematic sampling.

C. Experiment III

In this experiment, we are taking 2 different subjects - Subject 1 and Subject 2. We also take 1 sample and its respective subsamples from each of the subjects. Table 3 shows the ICP registration error between sample 1 along with its respective subsamples from subject 1 and sample 1 along with its respective subsamples from subject 2.

Random						
Table 3		**Subject 2**				
		Sample 1	Subsample 11	Subsample 12	Subsample 13	MSE
Subject 1	Sample 1	17.9131	17.8738	17.8848	17.8028	0.0602317
	Subsample 11	17.934	17.895	17.9056	17.8241	0.0467484
	Subsample 12	17.9335	17.8947	17.9052	17.8238	0.0468818
	Subsample 13	17.9337	17.8949	17.9051	17.8236	0.0469836
	MSE	0.017869876	0.025208629	0.015680800	0.094962926	

Systematic						
Table 3		**Subject 2**				
		Sample 1	Subsample 11	Subsample 12	Subsample 13	MSE
Subject 1	Sample 1	17.9131	17.863	17.9607	17.9561	0.0406963
	Subsample 11	17.9331	17.8829	17.9807	17.9761	0.0496255
	Subsample 12	17.9321	17.882	19.9799	17.9752	1.0340269
	Subsample 13	17.9317	17.8819	17.9792	17.9743	0.0485645
	MSE	0.016635504	0.036615229	1.034754436	0.0579220381	

Stratified						
Table 3		**Subject 2**				
		Sample 1	Subsample 11	Subsample 12	Subsample 13	MSE
Subject 1	Sample 1	17.9131	17.9367	17.905	17.9156	0.0125381
	Subsample 11	17.935	17.9588	17.9276	17.9373	0.0289999
	Subsample 12	17.9326	17.9564	17.9245	17.9351	0.0267819
	Subsample 13	17.9338	17.9582	17.926	17.9363	0.0281387
	MSE	0.017946796	0.040479470	0.0119606647	0.0200881183	

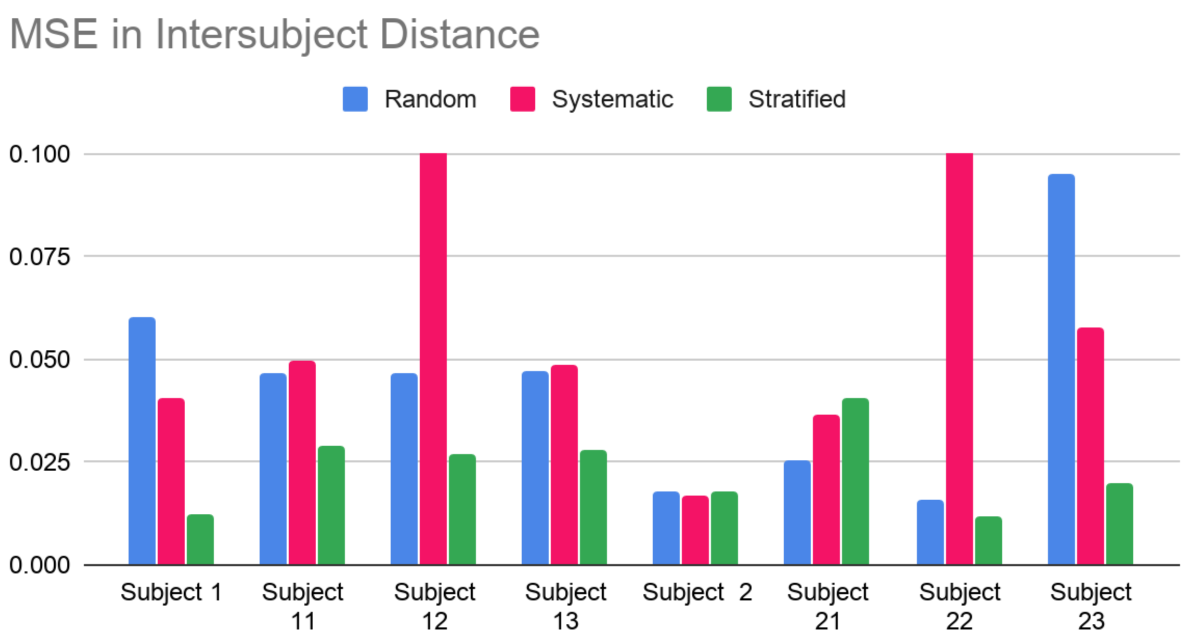

Table 3

INFERENCE: From the above experiment, we can see that the registration error using the subsamples is very similar to that using the original samples. From the graph, it can be inferred that the stratified sampling gives the least errors when comparing intersubject distance.

D. Experiment IV

For 5 samples each from a different subject, we create 30 subsamples each (to apply Central Limit Theorem). According to CLT the average mean and the average S.D. of the samples is similar to the mean and S.D. for the original population. Table 4.1 and Table 4.2 verify the CLT on the original samples and their subsamples.

CLT- Mean					Error		
Table 4.1	Original	Random	Systematic	Stratified	Random	Systematic	Stratified
Sample 1	88.4534246	88.4179602	88.4705998	88.4427378	0.03546440	0.01717525	0.01068678
Sample 2	81.8630944	81.8318121	81.8680909	81.8614916	0.03128226	0.00499654	0.00160273
Sample 3	79.6070373	79.5912596	79.5992354	79.6156961	0.01577766	0.00780187	0.00865886
Sample 4	94.3822379	94.3437242	94.4160172	94.3711543	0.03851362	0.03377930	0.01108353
Sample 5	95.8262718	95.8264071	95.8986898	95.8534787	0.00013531	0.07241801	0.02720693

CLT- SD					Error		
Table 4.2	Original	Random	Systematic	Stratified	Random	Systematic	Stratified
Sample 1	70.0752603	70.0866235	70.1114493	70.0821552	0.01136322	0.03618895	0.00689486
Sample 2	53.8800671	53.8601913	53.8651443	53.8841712	0.01987577	0.01492285	0.00410412
Sample 3	60.7547633	60.7816160	60.7837948	60.7658890	0.02685266	0.02903143	0.01112567
Sample 4	61.5487277	61.5256783	61.5333663	61.5484533	0.02304947	0.01536147	0.00027448
Sample 5	55.9003668	55.8964762	55.9162529	55.8848509	0.00389063	0.01588605	0.01551588

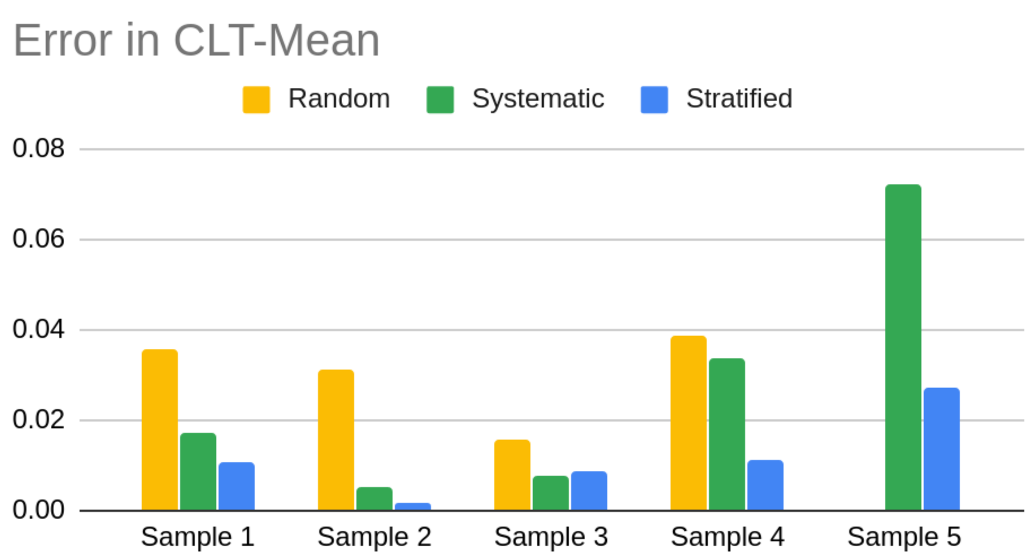

Table 4.1

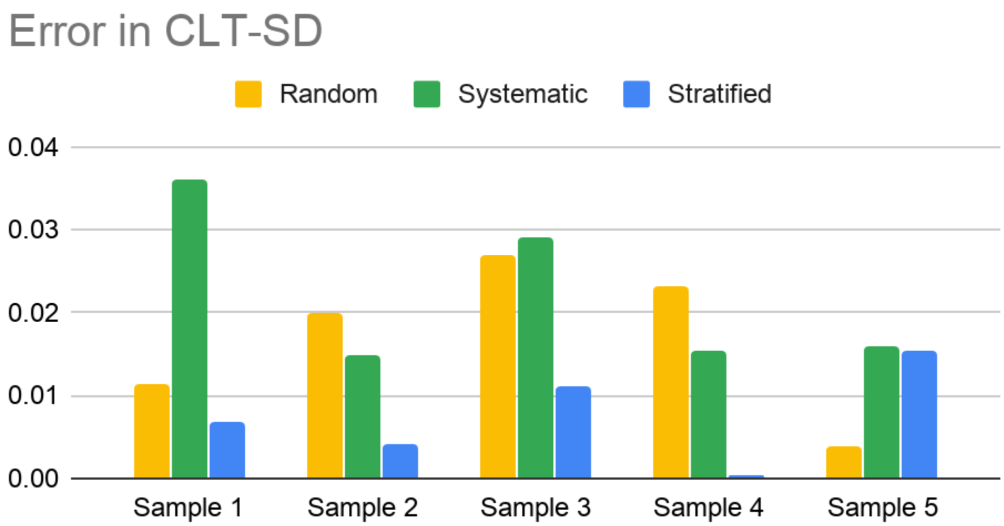

Table 4.2

INFERENCE: Results from the above experiment prove that the sub-samples created from the original sample have the same discriminative power as the original sample. From the graphs, we can infer that the stratified sampling is the best technique as it has the least errors in mean and standard deviation for all samples.

Conclusion

In this project, we used three different sampling techniques and proved that the sub-samples created using these techniques all carry the same information and have the same discriminative power as the original sample.

Following table ranks these three techniques in different criteria. From the table, it can be seen that the stratified sampling technique is the best overall.

Technique	Computational Time	Subsample Similarity	Sample-Subsample Similarity	Intersample Difference	Intersubject Difference	CLT
Random	1	1	3	3	2	3
Systematic	3	3	2	2	3	2
Stratified	2	2	1	1	1	1