Statistical Analysis: Fault-Tolerant Message Streaming

Problem Statement: Company X does fault-tolerant message streaming work. Specifically, you guarantee your customers 99% success of messages delivered. Recently, your company lost a few servers but your bosses are convinced that things are still working fine since things still seem to be working fine, So our task for this assignment was to statistically analyze the datasets from pre & post removal of servers and help the company gauge impact of their service.

Introduction: So for this problem at hand we will try to analyze the performance of the message streaming service with the help of statistical methods & visualize the results for business case description and understanding, To begin with, let us assume we have the following data sets:

1) X1, Let's assume that this is an old dataset pre-removal of servers.
2) X2, Let's assume that this is a new dataset post removal of servers.

Data-Sets/Samples: For this problem, we as a statistical team have been given access to some high-level data of messaging services. So our X1 & X2 data samples mentioned in the introduction section are composed of the following fields:

1) Message - This field contains details about the message(mostly numeric for analysis purposes).
2) Delivery_status - This field contains binary information of message delivery status :
- 0 - Failed to deliver
- 1 - Successfully delivered.

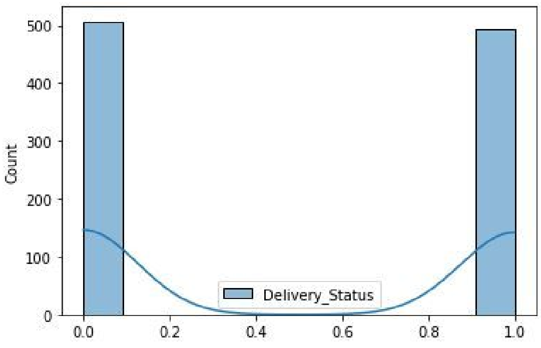
Our data sample of X1 & X2 has SIZE=1000

Method & Statistics:

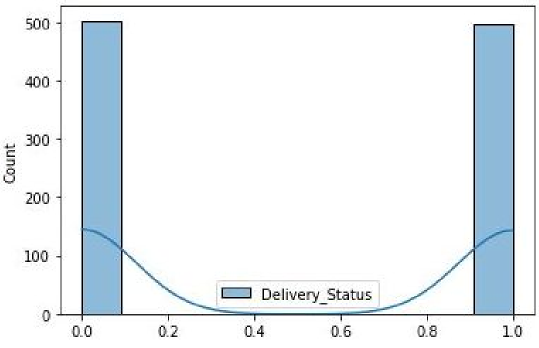
To begin with our analysis on a high-level, we try to analyze the "Delivery-Status" of message pre & post removal of servers, on our data samples X1 & X2 of size=1000 respectively.

Having a closer look at this binary field looks like Bernoulli distribution, where {0 & 1} value is experiments repeated multiple times.

So X1[pre-server removal] Delivery status Vs X2[pre-server removal] Figure Below:



VS



X1- Delivery statuses

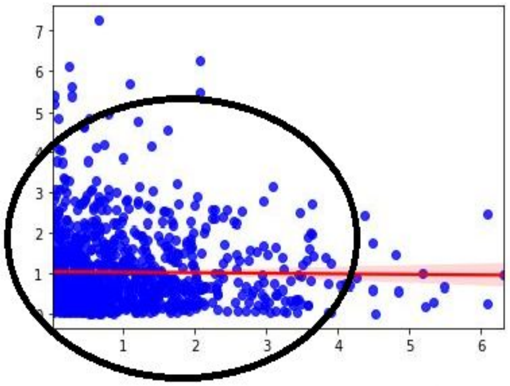
X2- Delivery statuses

So looking at the sample of size 100 X1 vs X2 we conclude that removal of servers for message streaming is not impacting the delivery status or degrading the companies message streaming performance.

Statistical Method:

After having an overview of messaging systems performance based on pre & post server removal, let us now deep dive and have a statistical method to have a look at it.

We used the confidence interval approach to see if our datasets {X1(pre) & X2(post)} has any overlap between their 95% range groups, for better understanding we visualize these confidence intervals using regplot, please refer figure below:

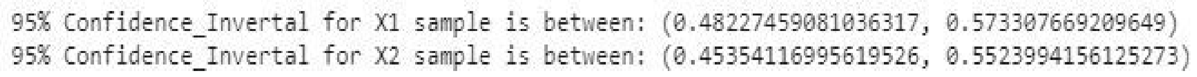


By Visualizing the {X1 & X2} with Confidential Interval regplot of 95% confidence we see a lot of datapoints are overlapped & large concentration of points can be seen on left corner, Proves that there is no statistical change in X1(old) vs X2(new) as well.

As we see a lot of concentration of point on left corner for confidence interval groups of X1 & X2, we can safely conclude by our statistical analysis that there is no difference in the performance of message streaming service after removal of servers.

Statistical Method 2:

Above C.I method worked upon samples {X1 & X2} and we see no difference visually/statistically, however, it is always advised to try multiple samples to conclude something for the large datasets, hence we would take the approach of mean & standard deviation of dataset & try finding the confidence to bolster out first 2 approaches & conclude the result.



```
95% Confidence_Invertal for X1 sample is between: (0.48227459081036317, 0.573307669209649)
95% Confidence_Invertal for X2 sample is between: (0.45354116995619526, 0.5523994156125273)
```

As from the above figure, we see there is a lot of overlap between the 95% confidence interval, hence it would be safe to conclude from our statistical analysis that, Statistically there is no significant difference.

Conclusion: Hence by the above methods & experiments, we can say there is no significant degradation in message streaming performance services of company X.

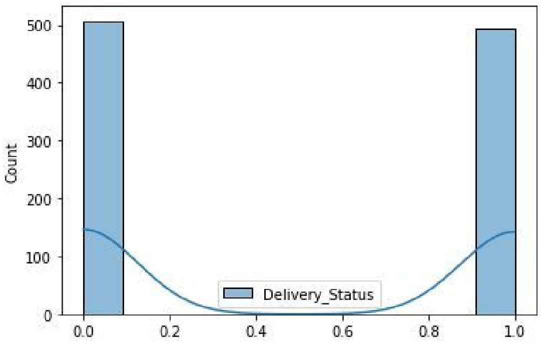
Business Justification: Fault-Tolerant Message Streaming

Problem Statement: Company X does fault-tolerant message streaming work. Specifically, you guarantee your customers 99% success of messages delivered. Recently, your company lost a few servers but your bosses are convinced that things are still working fine since things still seem to be working fine, So this report is to publish the result for business justification on performance degrade of message stream services, due to lost servers.

Introduction: By running experiments on the datasets of pre & post server reduction we tried to find the performance difference in both scenarios, Further detailed analysis and conclusions will be shared below.

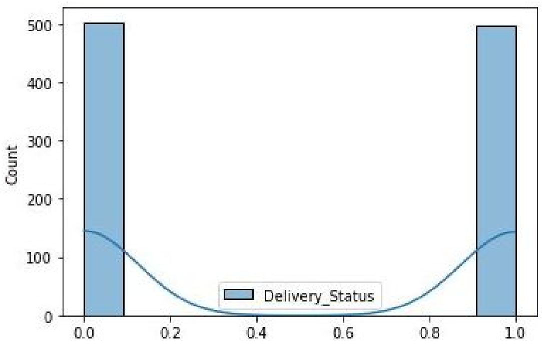
Overview:

We compared X1- Old(pre-server removal) vs X2 - New(Post-server removal) to see the degradation in the performance of the message streaming service.



X1- Delivery statuses

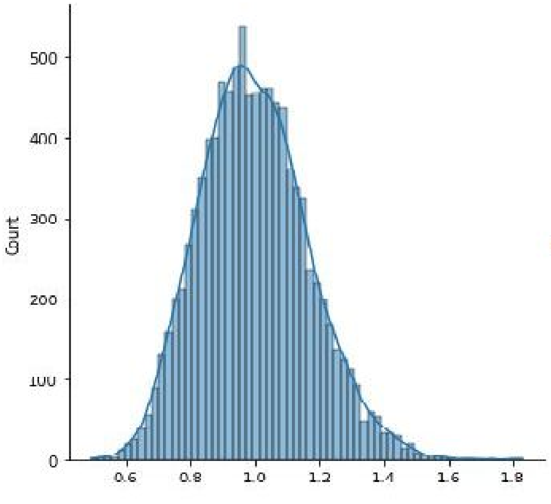
VS



X2- Delivery statuses

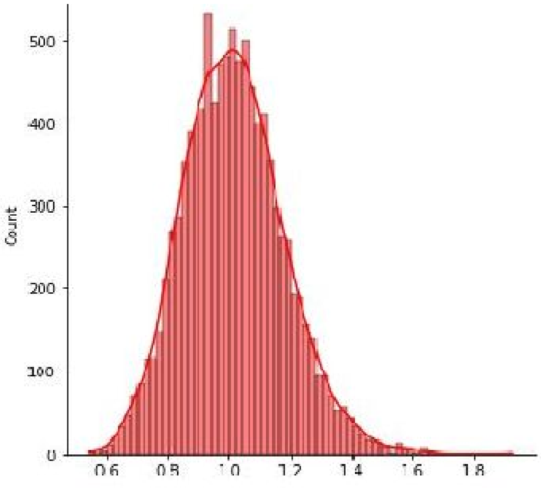
For the selected sample randomly we see 50% is the success rate of the messages, for both scenarios.

Both data sets have the same shape & follow the same distribution.



X1 Distribution

X1 Vs X2



X2 Distribution

As from the figure above, it is both data sample X1 & X2 follows the same pattern & shape, Hence we can conclude that there is no significant difference.

Conclusion:

From the above visualization & experiment setup, we can safely conclude that there is no significant degradation in the message streaming services of company X after losing servers.