Picking A Least Biased Random Sample Of Size n From A Data Set of N Points {Version 3}. ISSN 1751-3030

Abstract
In this research investigation, a Statistical Algorithm is detailed that enables us to pick a Least Biased Random Sample of Size n, from a Data Set of N Points with n<N

Article body
Picking A Least Biased Random Sample Of Size \( n \) From A Data Set of \( N \) Points \{Version 3\}  

ISSN 1751 - 3030

Author: 
Ramesh Chandra Bagadi  
Date Scientist  
International School Of Engineering (IN80FE)  
Postal Address: Plot No 63/A, 1st Floor,  
Road No 13, Film Nagar, Jubilee Hills,  
Hyderabad – 500033, Telengana State, India.  
Email: ramesh.bagadi@insoft.edu.in

Abstract  
In this research investigation, a Statistical Algorithm is detailed that enables us to pick a Least Biased Random Sample of Size \( n \), from a Data Set of \( N \) Points.  

Theory  
Given a Data Set of \( N \) points, if we were to pick a Least Biased Random Sample of Size \( n \), i.e., \( n \) Data Points, we can use the following stated Algorithm.  

Algorithm  
Firstly, we consider all possible Partitions of Size \( n \) of the given Data Set of \( N \) points. These will be \( ^N\!C_n = \frac{N!}{n!(N-n)!} \) in number. Let these be represented by \( P_i \) for \( i = 1 \) to \( ^N\!C_n \).  

Now, for such Partitions \( P_i \), we find the Average (Arithmetic Mean) \( \bar{X}_{P_i} \).  

We now find, using K-Means Clustering Algorithm, \( n \) Clusters using these \( ^N\!C_n \) data points called \( \bar{X}_{P_i} \) and find their Centroids and let us Label these \( \bar{A}_{P_i} \).
We now pick any particular Partition, say $P_k$, wherein we establish $n!$ Number of One-One Functions between the $n$ Elements of $P_k$ and the aforementioned $n$ Elements of Set $\overline{A}_{P_i}$ and Pick One that Particular Function such that the

a. Differences $|\overline{A}_{P_i}(l) - P_k(m)|$ are Minimum Possible for $l = 1$ to $^NC_n$ and $m = 1$ to $^NC_n$.

b. Sum Of the Differences $\sum_{i=1}^{^NC_n} |\overline{A}_{P_i}(l) - P_k(m)|$ are Minimum Possible for $j = 1$ to $^NC_n$ and $m = 1$ to $^NC_n$.

c. Sum Of the Squares Of the Differences $\sum_{i=1}^{^NC_n} (|\overline{A}_{P_i}(l) - P_k(m)|)^2$ are Minimum Possible for $j = 1$ to $^NC_n$ and $m = 1$ to $^NC_n$.

We now repeat this procedure for all the rest of $P_i$ other than $P_k$ and whichever Partition has this Least value, we consider that particular Partition has the Least Possible Sampling Bias.

**Finding the aforementioned $n!$ Number of Functions**

Considering the Set $\overline{A}_{P_i}$, the elements of the Set $P_k$ can be arranged among themselves in $n!$ Number of ways. Now the One-One position wise respective correspondence between the Elements of the Set $\overline{A}_{P_i}$ and the Elements of each of the aforementioned arrangements of the Set $P_k$ gives us the $n!$ Number of Functions.

We can also repeat the same Procedure using the Expected Value in place of the Mean $\overline{A}_{P_i}$. 
References


http://vixra.org/author/ramesh_chandra_bagadi


Medical Dictionary - 'Sampling Bias' Retrieved on September 23, 2009


