Semester : 4 Subject : Statistics for AI&DS Academic Year: 2023 2024.

BIAS:

* Statistical Bias refers to measurement or sampling errors that are systhematic and produced by the measurement or sampling process.

* There are errors due to random chance and errors due to bias.

* An unbiased process will produce error, but it is random and does not tend strongly in any direction. (Refer PPT for diagram).

* In biased process there is error and it is biaied also.

SIZE V/s QUALITY:

* When the data is huge, we naturally choose random sampling. Time and effort spent on random sampling reduces bias and also one can pay greater attention to data exploration and data quality.

* In millions of records, it is not feasible to consider missing data and outliers. Sometimes they contain useful and important information.

* But if there is a random sample of some thousand records, then it is feasible to track down.

* Let us consider the search queries from Google, here coloumns are terms, rows are individual search

queries and cellvalues are either 0 to 1.

* The aim is to determine the best predicted search destination for a query. There are over 1,50,000 words in the English language, and Google processes over one trillion queries per year. This is a true big data problem. But by popular search term, this is not at all a big problem.

* The value of search technology lies in the ability to return useful and detailed results for a huge variety of search queries.