(a) (2 points) Describe the advantage and disadvantage of using sampling to reduce the number of data points.

(b) (5 points) What is simple random sampling (without replacement)? Explain when we should use simple random sampling and when we shouldn't use simple random sampling. Give an example for each situation.

(c) (3 points) Describe what is noise and what is outlier? Which one is acceptable?

Ans.

a)

|  |  |  |
| --- | --- | --- |
| **Sr No.** | **Advantages** | **Disadvantages** |
| **1** | Smaller data size allows the use of a more expensive algorithm. | There can be a bias or error in collecting data |
| **2** | Low cost for data collection | Difficult to select sample that truly represents the data set. |
| **3** | Having correctly collected a sample, the accuracy of data is high. | Data collected from population may not be accurate. |
| **4** | Using a sample data set takes less time. | Using sampling requires knowledge of statistical analysis and probability. Lack of this knowledge may result in inaccurate. |

b) Simple random sampling without replacement is the process of randomly selecting an item from data population for the sample and removing it from the population.

Example: We have a stock of certain item and we want to check the quality of the batch. For this, we will use simple random sampling, as it is very difficult for our stock to check each and every item.

Simple random sampling should be used when the data population size is very large, as it will help in faster process.

Eg. Using population census data to find average age where the population size is in billions and it is very difficult to calculate using each value. So, we will use simple random sampling and take fewer values, like a few hundred thousand, for calculations.

Simple random sampling shouldn’t be used when the population is varied as it can fail when to create a proper sample set representing a population.

Eg. When building classification models for rare classes, it is critical that the rare classes be adequately represented in the sample. Using population census data to find state wide population density

c) Outliers are data objects whose characters are significantly different from rest of the data.

Noise is the modification of original value. It may involve the distortion of a value or the addition of spurious data.

Outliers are legitimate data objects and hence may be of use. Hence outliers are acceptable.