Statistical Data Mining

Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas. It is concerned with the analysis of large databases in order to find unsuspected statistical relationships which are of interest or value to decision makers.

The term data mining is not new to statisticians. It has a derogatory connotation because a sufficiently exhaustive search will certainly throw up new patterns because of the data are not simply uniform but have differences which can be interpreted as different patterns.

The problem is that many of these “patterns” will simply be a product of random fluctuations. The object of data analysis is not to model the new random patterns, but to model the underlying structures which give rise to consistent and recurrent patterns.

Data sets may be large because the number of observations is large or because the number of variables is large. When the number of variables is large the problem of dimensionality really begins to bite—with 1,000 binary variable which makes even a billion observations into insignificance

Statistics is a branch of mathematics that concerns the collection, organization, analysis, interpretation, and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied. Statistics is concerned with the use of data in the context of uncertainty and decision making in the face of uncertainty.

A **random variable** is described informally as a variable whose values depend on outcomes of a random phenomenon. In probability theory, a random variable can be defined as a measurable function defined on a probability space that maps from the sample space to the real numbers.

The domain of a random variable is called a *sample space,* defined as the set of possible outcomes of a non-deterministic event. For example, in the event of a coin toss, only two possible outcomes are possible: heads or tails.

Commonly, x is used to symbolized the random variable. When the range of x is countable, the random variable is called a **discrete random variable**. If the range is uncountable infinitely (usually an interval) then x is called a **continuous random variable**.

In statistics, a **population** is a set of similar items or events which is of interest for some question or experiment. A statistical population can be a group of existing objects or a hypothetical and potentially infinite group of objects conceived as a generalization from experience. A common aim of statistical analysis is to produce information about some chosen population.

In statistics a **sample** is a set of individuals or objects collected or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observations.

Typically, the population is very large, making a complete enumeration of all the individuals in the population impossible. The sample usually represents a subset of manageable size. Samples are collected and statistics are calculated from the samples.

The sample may be drawn from a population 'without replacement' (i.e. no element can be selected more than once in the same sample), in which case it is a subset of a population; or 'with replacement' (i.e. an element may appear multiple times in the one sample), in which case it is a multi-subset.

Type of samples

A **complete sample** is a set of objects from a parent population that includes all such objects that satisfy a set of well-defined selection criteria. Complete samples are such as the set of players in a major sports league, the birth dates of the master students, or a complete magnitude-limited list of astronomical objects.

An **unbiased (representative) sample** is a set of objects chosen from a complete sample, using a selection process that does not depend on the properties of the objects. An unbiased sample might consist of that fraction of a complete sample for which data are available, provided the data availability is not biased by individual source properties.

The best way to avoid a biased or unrepresentative sample is to select a random sample, also known as a probability sample.

A **random sample** is defined as a sample where each individual member of the population has a known, non-zero chance of being selected as part of the sample.

There are several types of random samples such as:

A **simple random sample** (or **srs**) is a subset of individuals (a sample) chosen from a larger set (a population) in which a subset of individuals are chosen randomly, all with the same probability.

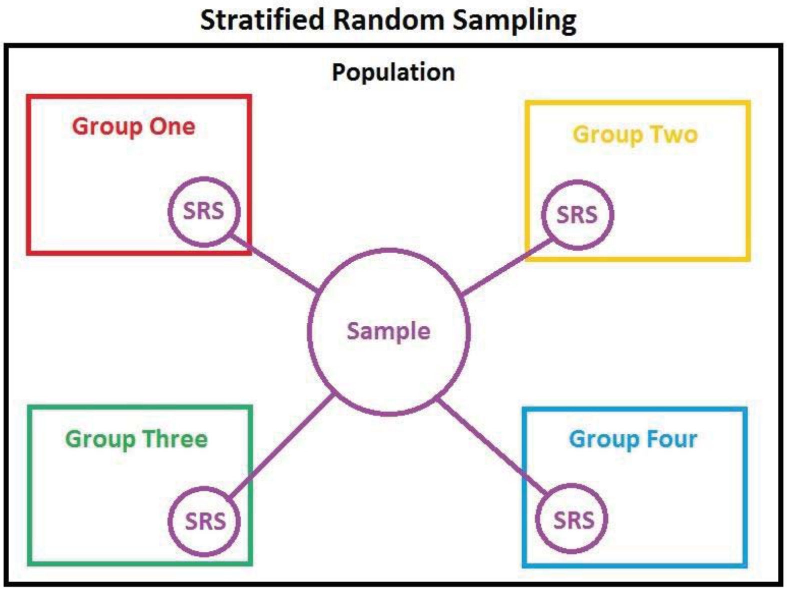
In srs, each subset of *k* individuals has the same probability of being chosen for the sample as any other subset of *k* individuals. A simple random sample is an unbiased sampling technique. Simple random sampling is a basic type of sampling and can be a component of other more complex sampling methods.

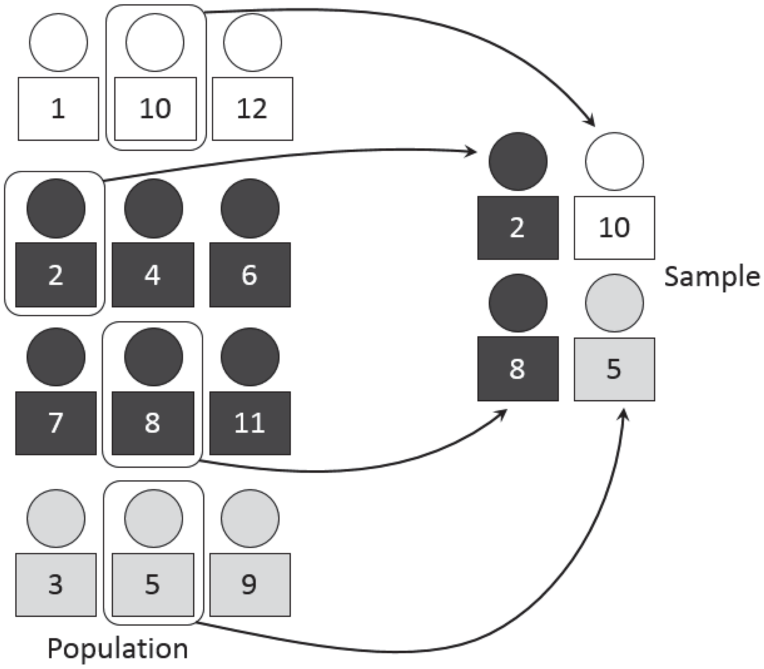
A **systematic sampling** is a statistical method involving the selection of elements from an ordered sampling frame. In this approach, progression through the list is treated circularly, with a return to the top once the end of the list is passed. The sampling starts by selecting an element from the list at random and then every k^{th} element in the frame is selected, where k, is the sampling interval (sometimes known as the *skip*): this is calculated as

$$k = N/n$$

where n is the sample size, and N is the population size.

A **stratified sampling** is a method of sampling from a population which can be partitioned into subpopulations. In statistical surveys, when subpopulations within an overall population vary, it could be advantageous to sample each subpopulation (**stratum**) independently. **Stratification** is the process of dividing members of the population into homogeneous subgroups before sampling. The strata should define a partition of the population. Every element in the population must be assigned to one and only one stratum. Then simple random sampling is applied within each stratum. The objective is to improve the precision of the sample by reducing sampling error.





A **cluster sampling** is a sampling plan used when mutually homogeneous yet internally heterogeneous groupings are evident in a statistical population. In this sampling plan, the total population is divided into these groups (known as clusters) and a simple random sample of the groups is selected. The elements in each cluster are then sampled. A common motivation for cluster sampling is to reduce the total number of interviews and costs given the desired accuracy.

A sample that is not random is called a non-random sample or a non-probability sampling.

sampling bias is a bias in which a sample is collected in such a way that some members of the intended population have a lower or higher sampling probability than others. It results in a **biased sample** of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling.

Medical sources sometimes refer to sampling bias as **ascertainment bias**. Ascertainment bias has basically the same definition, but is still sometimes classified as a separate type of bias.