e-PGPathshala
Subject : Computer Science
Paper: Data Analytics
Module 6: Sampling for Analytics
Module No: CS/DA/6
Quadrant 1 – e-text

1. Introduction

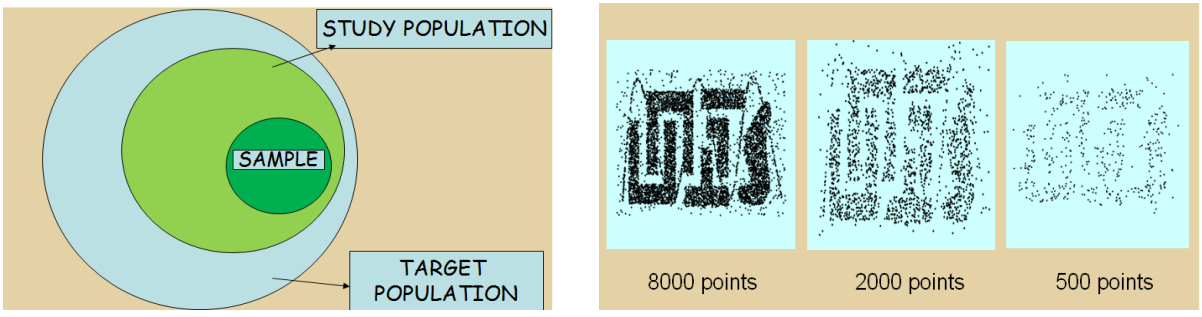
Sampling allows data scientists and data analysts to work with a small, manageable amount of data. So they can build and run analytical models more quickly and can produce accurate findings. When data sets are too large sampling is very useful for effective and productive analysis. Sampling is very useful in big data analytics. An important issue is the size of the required data sample. Here we cover some sampling techniques, significance of resampling and some resampling methods.

1.1 Learning Outcomes

- To understand the basis of sampling
- To learn about various sampling techniques for analytics
- To know the importance of sampling

1.2 Sample

A sample is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” (Field, 2005). A **sample** includes one or more observations of data from the entire population. Depending on the sampling method, a sample can have lesser number of observations than the population. When doing research, it is very difficult to survey every item of a particular population because there may be a large number of item.



(a) (b)

Figure 1(a). Diagrammatic representation of Population vs Sample and 1(b) sample with of different sizes

A sample must be representative of the whole, highlighting the key features/properties of the population. From figure 1 we can find samples of different sizes and how they highlight the features of the whole.

1.2.1 Why do Sampling

Sampling is done because it is difficult to gather data from the entire population. Even if the population is small, the data may be needed urgently, and in the population the data collection may take too long even by the participation of everybody .So the advantages of sampling may be summarised as:

- Resources like time, money and manpower can be reduced
- Accurate results can be calculated mathematically

Three factors that influence sample representativeness are: Sample size, Participation (response) and Sampling procedure. The cases where we might need to sample the entire population are:

- When the population is very small
- When a large number of resources are available
- When expected responses are less

1.2.2. Types Of Sampling

There are two main types of Samples:

- Probability
- Non-probability

A probability sample is one in which each person in the population has an equal or minimum known chance (probability) of being selected, but in a non-probability sample some people have more, but unknown, chance than others of selection. Probability samples are based on the mathematical theory of probability. An assured way of providing equal probability of selection is to use the principle of random selection. For analytic processes probability sample is suitable.

Methods of probability and nonprobability sampling

i. Simple Random Sample

In statistics, a **simple random sample** is a subset of individuals (a sample) chosen from a larger set (a population). Each distinctive item is chosen randomly and entirely depends on chance. So each individual has the same probability for being chosen at any time during the sampling process .Each subset of *k* individuals has the same probability of being chosen for the sample as any other subset

of *k* individuals. This method is known as **simple random sampling.** A simple random sample is an impartial surveying technique.

Simple random sampling is a basic type of sampling, because it can be used for other more complex sampling methods. The principle of simple random sampling is that every object has the same probability of being chosen.

ii. Systematic Sampling

Another statistical sampling method is **systematic sampling**, which demands selection of elements from an ordered sampling.

Figure 3. Systematic Sampling

The common form of systematic sampling is an equal-probability method. Here, progression through the list is used circularly, with a return to the top once the end of the list is passed. An element will be selected in random from the list and then every k^{th} element in the frame is selected, this is the way by the sampling goes. Here *k* is the sampling interval or *skip,* this is calculated as:

K=N/n, where *n* is the sample size, and *N* is the population size.

iii Stratified Sampling

In stratified sampling, the entire population will be divided into different groups, known as strata. Then, a probability (usually a simple random sample) is derived from each group. It is another popular probability sampling method.
Stratified sampling has many advantages comparing with simple random sampling. For example, using stratified sampling, it is possible to vary the sample size required to achieve a given precision. Or it may be possible to increase the precision with the same sample size.

iii. Cluster Sampling

Cluster sampling refers to a sampling method that has the following properties.

- The entire population can divide into a number of *N* groups, called **clusters**.
- Randomly selects *n* clusters from N to include in the sample.
- The number of observations within each cluster K_i is calculated, and $K = K_1 + K_2 + K_3 + ... + K_{N-1} + K_N$.
- Each element of the population can be assigned to only one, cluster.

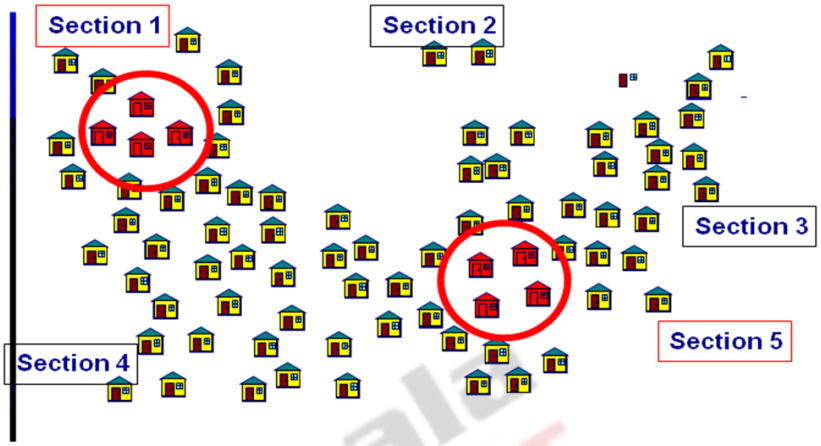


Figure 4. Clustered Sampling

Advantages and Disadvantages

The sample size is constant for cluster sampling methods. Cluster sampling generally offers less precision compared with simple random sampling or stratified sampling. One of the main disadvantage of cluster sampling is less precision.
Comparing with other sampling methods the sample point cost is less. So the researcher can manageable with a fixed budget to use a bigger sample with cluster sampling than with the other methods. When the sample size is increased it is sufficient to offset the loss in precision, such cases cluster sampling will be the best choice.

The Difference between strata and clusters

Strata and clusters are subsets of the population, but they differ in several ways.

- Every strata are represented in the sample, but only a subset of clusters are in the sample.
- Homogeneity of the elements within the strata will betterment the survey result by stratified sampling. Otherwise heterogeneity of the elements within the cluster will betterment the survey result by cluster sampling.

iv. Multistage Sampling

More complicated probability sampling method is multi-stage sampling. In multistage sampling larger clusters are again subdivided into more targeted smaller groups for

the purposes of surveying. Actually multi-stage sampling can be easier to implement and can create a representative sample of the concerned population. When a general sampling frame requires for preliminary construction, multi-stage sampling can help reduce costs of large-scale survey research. It also extent the aspects of a population which needs to include within the sample frame.

Multi-stage sampling begins first with the construction of the clusters. The first-stage clusters are further divided into second-stage cluster using a second element. Then identifies which elements to sample from which cluster until they are ready to survey.

Table 6(a) summary of the types and methods of probability sample.

Type of sample	**Selection strategy**
Simple random	Select from a full list of population called sampling frame. It can use a random number table to do this.
Systematic	Start random at any point on the sampling frame and choose every k^{th} case depending sampling size.
Stratified	Sampling frame are classified or stratified then apply random sampling.
Cluster	Population divided into units or clusters each containing individuals in a range of situations
Multistage	It is an extension of the cluster sample where samples are drawn from within clusters.

Table 6(b) Summary of the types and methods used in nonprobability sample.

Type of sample	**Selection strategy**
Convenience	Cases are selected depending on the availability.
Most similar	Select cases that are judged to represent similar or dissimilar cases, conditions or alternatively very different conditions.
Typical case	Cases selected that are known beforehand to be typical and not to be extreme.
Snowball	Group members identify additional members to be included in the sample.
Quata	Sample selected that yields the same

	proportions as the known population on easily identified variables.

1.3 Statistics

Statistics is the study of collection, analysis, clarification, presentation, and organization of data. To apply **statistics** in any type of problem, it is used to begin with a **statistical** population or **statistical** model process .

- Let S_i denote the random variable corresponding to data point x_i , then a statistic Ө is a function
  Ө : (S1, S2, · · · , Sn) → R.
- "a point estimate" of a parameter is the value of a statistic to estimate a population parameter, and the statistic is called an estimator of the parameter.[1]

1.3.1 Examples of sample statistics

A limited number of observations which selected from a population on a systematic or random basis, will provide generalizations about the population.

- Single population mean ,μ (known population standard deviation σ)
- Single population proportion, p
- Difference in means μ_1,μ_2 (t-test)
- Difference in proportions p_1,p_2 (Z-test)
- Odds ratio/risk ratio
- Correlation coefficient
- Regression coefficient

1.3.2 Frequency and Mode

Frequency is the number of occurrences of a repeating event per unit time. It is also called as **temporal frequency**, which highlights the difference between the spatial frequency and angular frequency. The **period** is the spell of time of one cycle in a repeating event, so the period is the reciprocal of the frequency.

The **mode** is the value that appears generally in a data set. The **mode** of a discrete probability distribution is the value p at which its probability mass function takes its maximum value, other words it is the value that is frequently sampled.

1.3.3.Percentile

A **percentile** or a centile is a measure used in statistics, which indicates a group of values fall below a given percentage of observations. For example, the 10^{th}

percentile is the value (or score) below which 10 percentage of the observations may be found.

The term percentile and the related term percentage are often used in the reporting of scores from NRT (norm referenced test). For example, if a score is in the 80th percentile, it is higher than 80% of the other scores. The twenty fifth percentile is known as the first quartile (Q_1), the fiftieth percentile as the median or second quartile (Q_2), and the seventy fifth percentile as the third quartile (Q_3). In general, percentiles and quartiles are specific types of quartiles.

1.3.4 Measures of Central Tendency: Mean and Median

“A measure of central tendency is a single value within a set of data that attempts to identifying the central position within that data. Measures of central tendency is sometimes referred as measures of central location. They are also termed as summary statistics. The mean (often called the average) ,meadian and mode are probably the measure of central tendency”.

Even though mean, median and mode are acceptable measures of central tendency, but in different conditions, some measures of central tendency become more appropriate to use than others. Next sections, we will look more about the mean, mode and median.

Mean

Most acceptable measure of central tendency is mean or average . It will go with both discrete and continuous data. Generally, it is using with continuous data. The mean is equal to the sum of all the values in the data set divided by the total number of values in the data set. So, if we have n values in a data set and they have values x_1, x_2, ..., x_n, the sample mean, usually denoted by $\bar{x}$ (sounded as x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

This formula is generally denoted as Greek capitol letter, Σ (pronounced "sigma"), which means "sum of...":

$$\bar{x} = \frac{\sum x}{n}$$

Median

The median is used to describe whole set of observations with a single value which represents the center of the data. It divides the data equally, that is half of the

observations are above the median and the next half are below it. It is determined by ranking the data and finding observation number [N + 1] / 2. If there are an even number of observations, the median is anticipated as te value halfway between that of observation numbers N / 2 and [N / 2] + 1.

7 9 10 12 **13** 14 17 18 19

In the above ordered data, the median is 13. That is, half percentage of the values are less than or equal to 13, and other half of the values are greater than or equal to 13.

1.3.5 Measures of Spread: Range and Variance

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population. It is usually used in conjunction with a measures of central tendency, such as the mean or median, to provide an overall description of a set of data.

Range

The range is the difference between the highest and lowest scores in a data set and is the simplest measure of spread. So we calculate range as:

Range = maximum value - minimum value

For example, let us consider the following data set:
23 56 45 65 59 55 62 54 85 25

The maximum value is 85 and the minimum value is 23. This results in a range of 62, which is 85 minus 23. Whilst using the range as a measure of spread is limited, it does set the boundaries of the scores. This can be useful if you are measuring a variable that has either a critical low or high threshold (or both) that should not be crossed.

Variance

In probability theory and statistics, **variance** measures how far a set of numbers are spread out. A variance of zero indicates that all the values are identical. Variance is always non-negative: a small variance indicates that the data points tend to be very close to the mean(expected value) and hence to each other, while a high variance indicates that the data points are very spread out around the mean and from each other.

An equivalent measure is the square root of the variance, called the standard deviation. The standard deviation has the same dimension as the data, and hence is comparable to deviations from the mean.

Example:

A random sample of 10 college students reported sleeping 7, 6, 8, 4, 2, 7, 6, 7, 6, 5 hours, respectively. What is the sample standard deviation?

The sample variance is:

$$\text{variance, } s^2 = 1/19[(7-5.8)^2+(6-5.8)^2+\cdots+(5-5.8)^2]=1/19(27.6)=3.067$$

Therefore, the sample standard deviation is:

$$\text{Standard deviation, } s = \sqrt{3.067} = 1.75$$

Thus in statistics, quality assurance, and survey methodology, sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population. Each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals. In survey sampling, weights can be applied to the data to adjust for the sample design, particularly stratified sampling. Results from probability theory and statistical theory are employed to guide the practice. In business and medical research, sampling is widely used for gathering information about a population.

Case Studies

Epidemiology: Methods of Sampling from a Population

Source: http://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population

It would normally be impractical to study a whole population, for example when doing a questionnaire survey. Sampling is a method that allows researchers to infer information about a population, without having to investigate every individual. Reducing the number of individuals in a study reduces the cost and workload, and may make it easier to obtain high quality information, but this has to be balanced against having a large enough sample size with enough power to detect a true association.

If a sample is to be used, by whatever method it is chosen, it is important that the individuals chosen are representative of the whole population. This may involve specifically targeting hard to reach groups. For example, if the electoral roll for a town was used to identify participants, some people such as the homeless would not be registered and therefore excluded from the study by default.

There are several different sampling techniques available.

1. Simple random sampling

In this case each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected. One way of obtaining a random sample is to give each individual in a population a number, and then use a table of random numbers to decide which individuals to include.

2. Systematic sampling

Individuals are selected at regular intervals from a list of the whole population. The intervals are chosen to ensure an adequate sample size. For example, every 10th member of the population is included. This is often convenient and easy to use, although it may also lead to bias for reasons outlined below.

3. Stratified sampling

In this method, the population is first divided into sub-groups (or strata) who all share a similar characteristic. It is used when we might reasonably expect the measurement of interest to vary between the different sub-groups. Gender or smoking habits would be examples of strata. The study sample is then obtained by taking samples from each stratum.

In a stratified sample, the probability of an individual being included varies according to known characteristics, such as gender, and the aim is to ensure that all sub-groups of the population that might be of relevance to the study are adequately represented. The fact that the sample was stratified should be taken into account at the analysis stage.

4. Clustered sampling

In a clustered sample, sub-groups of the population are used as the sampling unit, rather than individuals. The population is divided into sub-groups, known as clusters, and a selection of these are randomly selected to be included in the study. All members of the cluster are then included in the study. Clustering should be taken into account in the analysis. The General Household survey, which is undertaken annually in England, is a good example of a cluster sample.

5. Quota sampling

This method of sampling is often used by market researchers. Interviewers are given a quota of subjects of a specified type to attempt to recruit. For example, an interviewer might be told to go out and select 20 adult men and 20 adult women, 10

teenage girls and 10 teenage boys so that they could interview them about their television viewing. There are several flaws with this method, but most importantly it is not truly random.

6. Convenience sampling

Convenience sampling is perhaps the easiest method of sampling, because participants are selected in the most convenient way, and are often allowed to chose or volunteer to take part. Good results can be obtained, but the data set may be seriously biased, because those who volunteer to take part may be different from those who choose not to.

7. Snowball sampling

This method is commonly used in social sciences when investigating hard to reach groups. Existing subjects are asked to nominate further subjects known to them, so the sample increases in size like a rolling snowball. For example, when carrying out a survey of risk behaviours amongst intravenous drug users, participants may be asked to nominate other users to be interviewed.

Summary

- Sampling becomes very essential for analytics
- Basics of sampling is necessary before doing the real analysis