Sampling
Introduction to Data Science

Team 15:
Thijs van der Knaap (s2752077)
Hatim Alsayahani (s3183696)
Sebastian Wehkamp (s2589907)
Dimitris Laskaratos (s3463702)

September 24, 2017

Contents

1 The aim of this document

This document provides insight in several known concepts associated with sampling, as well as introducing and explaining new concepts. Most importantly, we will illustrate in what ways sampling is being used in science to understand data. Together with this document a Jupyter Notebook with examples of sampling techniques is also included.

First we will demonstrate a couple of traditional sampling techniques which are still widely used. Secondly there is a section about when you do sampling. Thirdly a step-by-step guide is included of how to actually apply sampling. Afterwards some advanced sampling techniques are discussed and lastly four non trivial questions are answered.

2 Traditional Statistical Sampling

Traditional Statistical Sampling refers to the selection of a small number of individuals from the statistical population in order to study a specific characteristic of the population. The selection is arbitrary, meaning, the individuals are not selected based on certain features or properties, but rather in a random manner.

A typical example could include the selection of 50 individuals of a town of 1000 people in order to approximate the median income of said town.

3 General-Purpose Sampling Strategies

The following strategies are heavily employed in statistics and surveys, to provide a general picture of certain aspects of the population:

3.1 Simple Random

Simple Random Sampling is the basic sampling method used for statistical analysis. It is applied to a population where the individuals are similar to one another on important variables. The person performing the sampling aims to acquire a set of individuals whose one more characteristics coincide, so that the analysis can be conducted based on them. In probabilistic terms, each individual of the population has the same chance of being selected.

The Simple Random method is prone to produce errors due to the randomness of the selection. The sample may not reflect the actual makeup of the population, something which other methods aim to overcome. Furthermore, this method is time consuming an tedious.

3.2 Systematic Sampling

In **Systematic Sampling**, the population under study is arranged in an ordering scheme. Subsequently, the individual selection is done at regular intervals (determined by the statistician), with the starting point chosen randomly from the list.The reason the starting point must be chosen randomly, as well as the usage of step, is to eliminate bias during the selection. However, Systematic Sampling has certain disadvantages, one of them being the fact that it is vulnerable to periodicity's which can lead to under-representation in the sample. Also its theoretical nature makes it difficult to measure its accuracy.

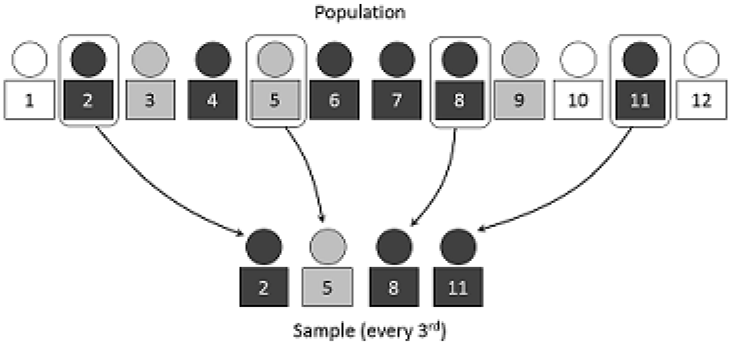


Figure 1: A visualization of Systematic sampling

3.3 Stratified Random Sampling

Stratified Random Sampling requires the statistical population to be divided into groups, called 'strata', based on a certain feature. The sampling process is then conducted on a group level, where the sample members are selected randomly from their corresponding groups. A requirement is that the statistical population must be heterogeneous in one or more attributes relevant to the research. This technique enables researchers to acquire information about the subgroups which could not be possible in a generalised random sample, while also providing more accurate statistical estimates. Lastly, since the subgroups are considered independent population the researcher may choose to apply different sampling techniques on each, according to his needs. However, other than the fact that stratified sampling is time-consuming and tedious, it can increase the cost and complexity of the selection. Furthermore, working with multiple criteria that are related to some, but not all, of the groups, may greatly impact the design of the research.

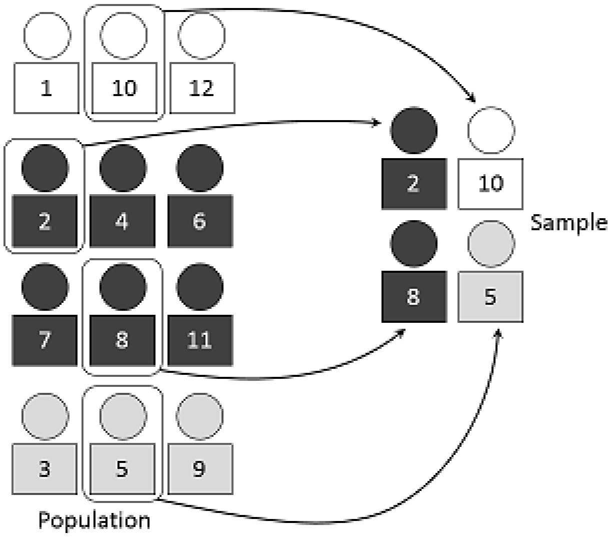


Figure 2: Stratified random sampling

4 Density-Biased Sampling

Analysing large data sets often requires a sampling of the data, in order to provide a general representation of them that can be processed more efficiently [1].

Density-Biased Sampling (DBS) is a special sampling technique, proposed for data mining that addresses the above problem. It is based on a probabilistic model which dictates that the probability a certain point will be included in the sample, depends on the density of the cluster it was selected from.

Another reason DBS is more widely used in data mining than uniform sampling, is because it bypasses the latter's severe drawback. That is, missing small clusters of data. An example is illustrated in Figure 3:

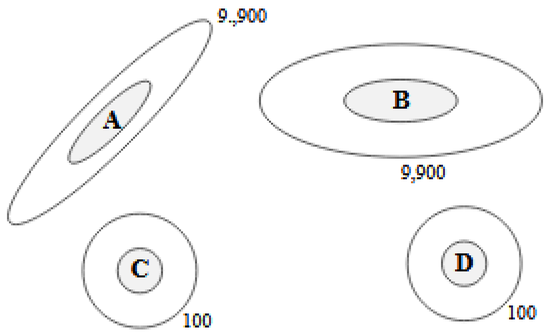


Figure 3: Four clusters with different data distribution

As you can see, the difference in the clusters is dramatic: clusters A and B contain 9.900 points each, while clusters C and D contain 100 each. If uniform sampling is applied here, is it expected that the vast majority of points included in the sample will be coming from clusters A and B, while any points selected from C and D will most likely be treated as noise by most clustering algorithms. On the contrary, by using Density-Biased Sampling, it is guaranteed that all clusters will have their corresponding representation in the sample, by taking into account the distribution of the points in them.

5 Embedded sampling

The easiest moment to apply sampling is by doing it in the pre-processing phase, you create your sample and run your program. As an alternative to this approach you can take the embedded approach. With embedded sampling, sampling becomes part of the program itself. An example of the embedded approach would be adaptive sampling. With adaptive sampling you start with a small sample and keep increasing it until the results are sufficient. While this technique eliminates the need to determine the correct sample size initially, it requires that there be a way to evaluate the sample to judge if it is large enough. This is for example the case for learning algorithms where you have a training data set.

A sampling strategy may produce better results when embedded in the procedure of an algorithm than if you would do the sampling first and then the algorithm. However only a couple of methods can be embedded and it may not be an easy task to incorporate sampling in your algorithm.

6 Step-by-step plan

At this point you are aware of which sampling techniques are out there and want to apply sampling to your own dataset. Where do you start? This section will contain the basic outline of which steps you have to go through in order to apply sampling.

1. The Type of data sets differ in many ways. For example data can be different types of quantative and qualitative data or it may have special characteristics; e.g. time series data. The type of data determines which sampling methods are better suited then others. An example would be time series data for which you could use Systematic sampling so you get a sample with an even distribution over time.

2. The second step is looking at the quality of the data set. This is often far from perfect and you want to cleanup your dataset. Examples of often encountered data quality issues are outliers; missing data of biased data. You have to clean this data up before you apply sampling. For reference take a look at one of the papers about missing values.

3. Check if you see something outstanding about the data. You could check if you can find e.g. groups in your original data set. This might make stratified random sampling a good solution. To do this you can make a number of plots of your data to easily see if there is something outstanding.

4. Determine your sample size. Before you apply sampling you have to think about how large you want your sample to be. There is no easy rule to know how large your sample should be since it can be completely different for every case. A more detailed description of how to determine this is given in subsection 12.2.

5. Determine whether it is possible and feasible to embed your chosen technique. Often embedding your sampling technique is possible with learning algorithms.

6. Apply your chosen sampling technique.

7 Bagging and boosting

Bagging and boosting are two methods that both use multiple weak learners to create one strong learner. A weak learner is a learner that has an correct classification rate of around 60%. Used individually these learners are useless due to there bad performance. Bagging or boosting ican be used when no strong learner is available.

7.1 Bagging

Bagging or **B**ootstrap **agg**regat**ing** creates for every weak learner a sample of the data. These subsets are uniformly sampled using replacements. Replacements means that data points aren't removed from the data set when they are added to a sample. A data point can therefore be present in several samples and appear multiple times in one sample.

Each sample is used to train one weak learner. All the trained weak learners are combined using majority voting or ? to create a strong learner that is capable of classifying data points with higher precision.

Bagging will reduce the variance and avoids over-fitting.

7.2 Boosting

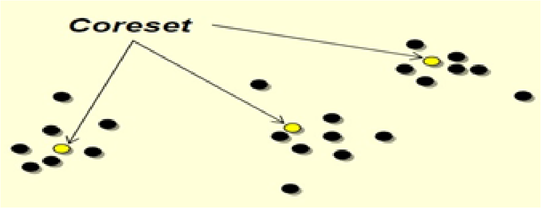
Boosting uses knowledge about the performance of each learner to achieve better results. Each weak learner is selected in such a way that it will correct for the miss-classifications by previous learners. Boosting uses at each iteration the whole training set. Boosting represents a group of strategies but most will follow this structure:

1. Set an equal weight for every element in the set.
2. while set of weak learners is not empty
   (a) Select the current best weak learner.
   (b) reduce the weight of the elements that are correctly classified and increase the weight of the wrong classified.
   #This allows the following selected weak learners to specialise in the wrongly classified elements.
3. Use your acquired strong learner on problems

Boosting will reduce variance and the bias.

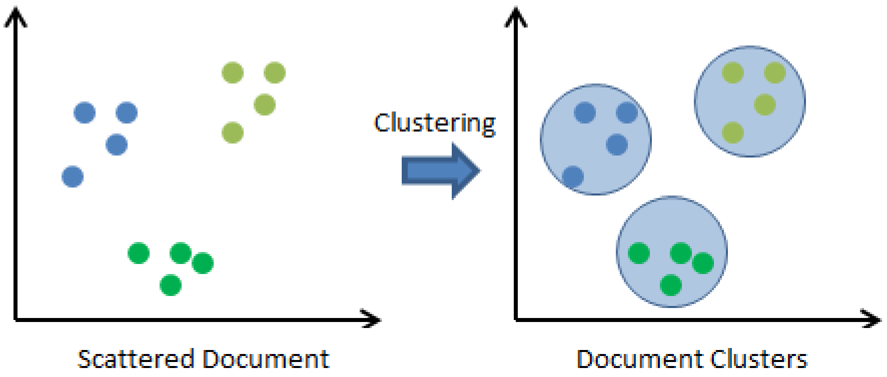
8 Representative Sample (Coresets)

Extracting and forming a small group from a large group which reflects and maintains the properties of the large group. For example, in Battalion a group of 10 soldiers, in which half of them are Lieutenants and half are Captains, representative sample could be done by extracting a small group of four soldiers two Lieutenants and two Captains.



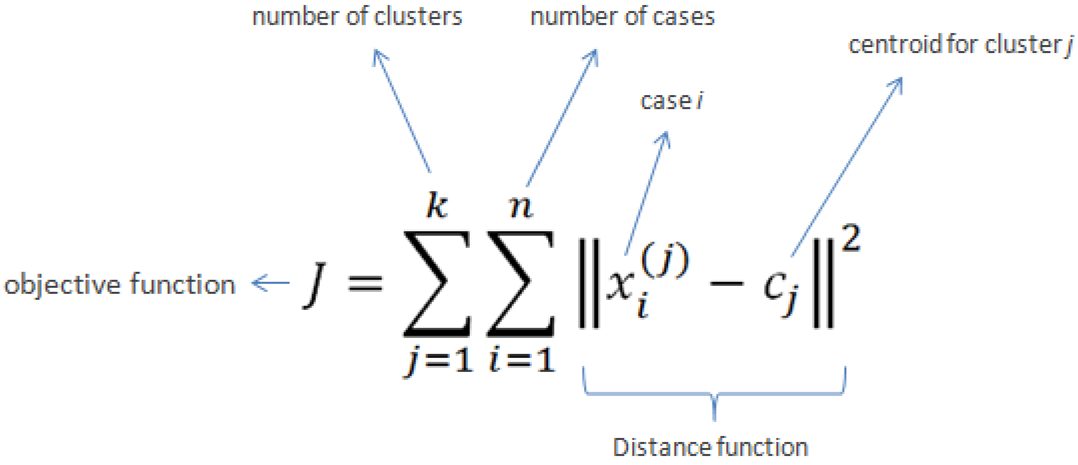
9 Clustering

dividing a group of data points into a number of groups according to their characteristics, it has more than 100 technique , in the following the most widely used clustering technique will be discussed.



9.1 k means clustering:

the iterative clustering algorithm that aims to grouping data into K groups. it is widely used for data mining and machine learning purposes.



$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

The illustration of k-means process

1. set the number of k.
2. measure the centroid point.
3. categorize each item to its closest centroid.
4. back to the second step because it is iteration process.

10 Sparse dictionary learning:

[2] modeling data vectors as sparse linear combinations of basis elements, to adapt it to specific data in image and eliminate some noise. (Machine Learning for Image Classification)

11 Summary

This paper discussed several popular sampling strategies like simple random sampling, systematic sampling, and stratified random sampling. For every sample strategy the way it works and a situation in which it is useful is given. section 5 briefly discusses the possibility of embedding sampling into your program and gives an example of a use case. Using the step-by-step plan of section 6 the reader should be able to choose the correct sampling method and implement it. Lastly several advanced strategies are discussed together with why they are useful.

12 Questions

12.1 Is Density-Biased Sampling a commonly used technique in statistics?

No. While most of the sampling techniques mentioned above are pretty much common practice in statistics in the macro-world, DBS is, almost exclusively, used in the micro-world. It's already been stated it can be used effectively in data mining, but another formidable example is its exploitation in studying pore formation in Biology[3]. DBS is basically a concept and, as such, its application in algorithms varies according to the researcher's needs.

12.2 How do you determine the required sample size?

1. You start by checking your population size. It is quite common for this to be estimated if you only have a sample.

2. A sample could always differ from the total population. The confidence interval shows how much higher or lower your results are compared to the population. An example would be "54% of the voters voted No with a margin of error of +-3%." This means that the results for the complete population are in the range of 51% and 57%.

3. How small do you want you confidence interval to be? This really depends on the use case but the most common intervals are 95% and 99%.

4. You have to determine your standard deviation. The most safe decision is to use .5 ensuring that your population is large enough

The chosen confidence interval corresponds to a fixed Z-Score. Below are the Z-scores of the most common ones, others can be looked up here.

$$95\% = 1.96$$

$$99\% = 2.576$$

You can now use the formula below

$$\text{Necessary Sample Size} = \frac{(\text{Z-score})^2 * \text{StdDev} * (1 - \text{StdDev})}{\text{CfdInt}^2}$$

12.3 When to use bagging or boosting?

For each you need multiple weak learners, so those have to be acquired. Use bagging when the classifiers are sensitive to the variations in the training set. Bagging is also used to prevent over-fitting. Boosting has many different implementations, but all suffer from noise sensitivity. This is because wrong classified elements will get more weight. So, use boosting when you have clean data and there is no danger of over-fitting, otherwise bagging might be a better choice.

12.4 why we use corsets in clustering and dictionary learning?

Fast,streaming,and parallel implementations One major advantage of corsets. Also reduce computational time and space complexity.

References

[1] Christos Faloutsos Christopher R. Palmer. Density biased sampling:an improved method for data mining and clustering. 1999.

[2] MichaFeigin DanFeldman and NirSo chen. Learning big(image)data via core sets for dictionaries. 2013.

[3] Michael Feig Vahid Mirjalili. Density-biased sampling:a robust computational method for studying pore formation in membranes. 2014.