Artificial Intelligence (AI) for Investments
Prof. Abhinava Tripathi
Department of Industrial & Management Engineering
Indian Institute of Technology – Kanpur

Lecture – 17
Statistical Inference

(Refer Slide Time: 00:13)

Welcome to the Statistical Inferencing, **(Video Starts: 00:17).** So, far, whatever you have learnt is more or less part of descriptive statistics. Using these we have been able to answer questions for the sample, but not really about the whole population. In fact, in most cases you do not have exact answers to the questions related to population majorly because the population is huge and collecting data for all of them is really, really difficult and cost intensive.

So, how do we proceed in such cases? We will try to discuss that some of that here. Let us start with a simple example from manufacturing industry that most of us might have heard off. A company that manufactures noodles is banned from producing noodle packets, with excessive amounts of lead. The quantity of lead that is banned is greater than 2.5 PPM, parts per million which is the accepted threshold in the industry.

Let us say that you are working as part of the food regulator company and you need to validate that the noodles do have large amounts of lead. Obviously, you cannot go to each and every factory and check each packet for the amount of lead, as it will take up too much time, money

and resources. There has to be some more efficient way. So, what if you do not check every packet, but instead take a small sample.

Find the average lead content in them and conclude the results for the whole bunch of noodle packets that are manufactured. Consider that you have around 30000 packets of noodles that this company has manufactured. The mean lead content of which is unknown to us. Out of these 30000 packets, we select around 100 packets randomly and measure the lead content in each of them.

And after this process we find out that the mean lead content in these 100 packets is 2.2 parts per million PPM, with a standard deviation of 0.7 PPM. Well, do you think these parameters would be same for the 30000 packets of population? How sure can you tell that it might have happened that the sample may be selected coincidentally had the packets which contained higher amounts of lead, or it could be the case that these packets had lower amounts of lead as well?

We just do not know that yet. One thing is for sure that we cannot at this point say anything decisively about the whole population of the noodle packets, but we will be able to by the end of this discussion. **(Video Ends: 02:36)**

(Refer Slide Time: 02:37)

Now, I would like to bring your focus to one key concept, which is sampling. **(Video Starts: 02:43) S**ampling is one of the key things that help us make better inferences. As you know, the crux of inferential statistics is that you need to infer the population from the parameters of the

sample and the way you create this sample plays a very important role. This is also used a lot while building machine learning models when you split them into train test and validate.

Remember that in our problem, as we mentioned earlier, we have to select 100 packets randomly suppose that we select these hundred randomly chosen packets from any one of the single factories. It could be possible that the issue with lead content was specific to that factory. Only and as a result, we end up banning the entire company. To avoid such problems food regulators need to ensure that the sampling procedure used is fair and unbiased.

So that later on no one can question the accuracy of the inference after the statistical analysis is done. **(Video Ends: 03:35)**

(Refer Slide Time: 03:35)

(Video Starts: 03:36) So, let us understand some of the different sampling methods that are commonly used in the industry. Let us take our noodles example only and see, what are the different ways in which we can select a sample that has 100 noodle packets? One way is that I get all the noodle packets from the population of 30000 and then just randomly select 100 packets. So, essentially, each packet has an equal probability of being selected.

This method of sampling is known as simple random sampling. One simple way of looking at simple random sampling is that you assign a number each element in the population and place all the numbers in a bag and then bring a blindfolded person who selects how many ever units you want to be present in your sample. You mostly use simple random sampling within a sampling method to generate the final sample.

Let us learn about these sampling methods. Suppose the company has two factories and we know that 70 percent of the noodles are manufactured in one factory A and 30 percent of the noodles in factory B. Now, you want to select 100 packets from our sample. So, we would select 70 packets from factory A using random sampling and 30 packets from factory B again using random sampling within the factory specific packets.

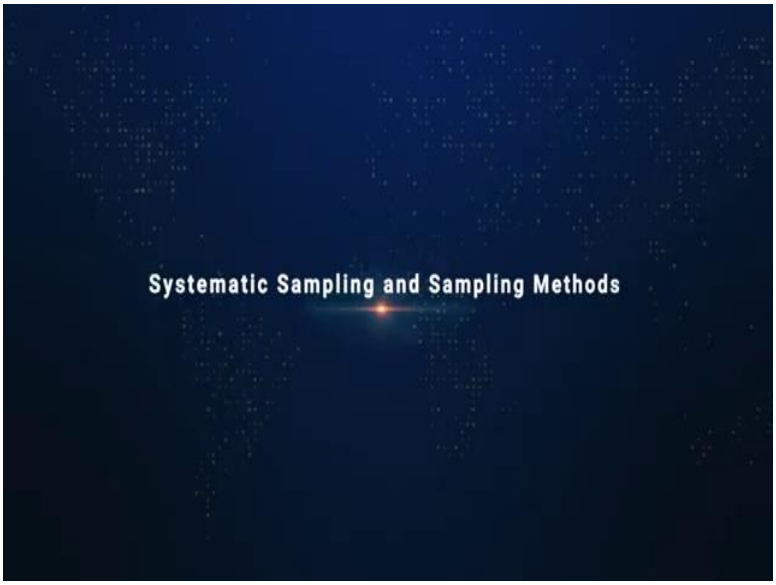
Thus now, again, you would have a sample of 100 packets, but this time your sample would be more representative of the population, as we have packets from both factories A and B and the proportions would be the same as that of the population. This approach is known as stratified random sampling. Here the units are divided into subgroups and then selected randomly.

But this is done in such a way that the final sample has the same proportion or approximately same of these subgroups, as does the population. Now, let us look at another approach. Let us say that you are collecting these noodle packets from warehouses across all the country. Let us suppose that there are around 20 warehouses in the country. So, what do you do in this approach is that you consider each warehouse as a cluster.

This means that there are 20 clusters in total. Now as part of your sampling, obviously, you will not analyse each cluster separately. Instead of these 20 clusters, you use simple random sampling and select any three clusters and consider all the packets within these 3 clusters. As your sample here you saw that we divided the population into clusters where each cluster is a warehouse.

And then we only selected a sample of these clusters using a simple random sampling. Since we are dividing the population into clusters, this sampling plan is known as cluster sampling. Cluster sampling is usually used when you see that the population can be divided into different groups or clusters that we have different characteristics. Once you create the clusters, you then use the similar ones or you choose dissimilar ones within which you can select all the members or do simple random sampling. **(Video Ends: 06:26)**

(Refer Slide Time: 06:26)



(Video Starts: 06:29) Now, let us say that in order to select your sample, what you did was to select every third packet, starting from the second packet on a list of the population. So, what did we do here? We selected a random starting point and started picking out sample units at some fixed or periodic interval. Such a kind of sampling is known as systematic sampling.

So, in this discussion we learned about four types of sampling methods, namely simple, random sampling, stratified random sampling, cluster sampling and systematic sampling. Which type of sampling techniques seems more suitable for choosing our noodle packets so that it represents our population? For many of the food regulators, stratified sampling is considered as one of the ideal sampling techniques that are usually most representative of the parent population.

However, in many cases, simple, random sampling is also used when the parent population is not heterogeneous in nature. What could be those cases where the parent population is not heterogeneous in nature? For instance, if all the noodle packets are of the same nature and manufactured in a single factory. Then simple, random sampling would be the most straightforward but if the packets came in different flavours and were manufactured in different factories.

Then, ideally stratified sampling would be the recommended method. One important thing to note is that all the sampling techniques learned until now are categorized under random sampling or probability sampling method. So, probability samples are types of samples when

every unit of the parent population has a known chance of being included in the sample. **(Video Ends: 08:04)**

(Refer Slide Time: 08:04)

(Video Starts: 08:08) Types of sampling, non-random sampling. There is another category of sampling when it comes to sampling methods and that is known as non-random sampling. Non-random sampling techniques are those where the odds of any sample unit being selected for a sample cannot be really calculated. We will discuss different non-random sampling methods. So, the first type of non-random sampling we look at is convenient sampling.

So, let us say you had to select 100 noodle packets in our previous example. And in order to do this, you simply choose the 100 packets that were around you and measure their lead content. As the name suggests, this sampling method is based on the convenience of the person who is selecting the sample. Clearly, this sampling method has the potential to become extremely biased.

The only good thing here is probably is that this is relatively convenient sampling method. The other type of sampling method that we will look at is called judgmental sampling. Again, as the name suggests, judgmental sampling is purely judgmental in which the sample units are chosen only on the basis of the knowledge and judgment of the person who is selecting the sample. Often the survey questions and responses require highly specialized skill set.

So, for example, if you wish to study the implication of blockchain on bank industry. Then it is advice that you will judgemental sampling to select people having relevant skill set to respond to the questions. **(Video Ends: 09:27)**

(Refer Slide Time: 09:27)

(Video Starts: 09:28) Let us, understand the difference between stratified sampling and cluster sampling. Now, some of us might get confused between stratified sampling and cluster sampling, as in both cases, we are dividing the population into subpopulations. In stratified sampling we divide the population into subpopulation and select the sample units in the same proportion as the subpopulations.

So, that the sample is as representative as the parent population. However, in the case of cluster sampling, we also divide the population into subpopulation or clusters, but, unlike stratified sampling, we only study selected clusters and not all the Clusters. **(Video Ends: 10:05)**