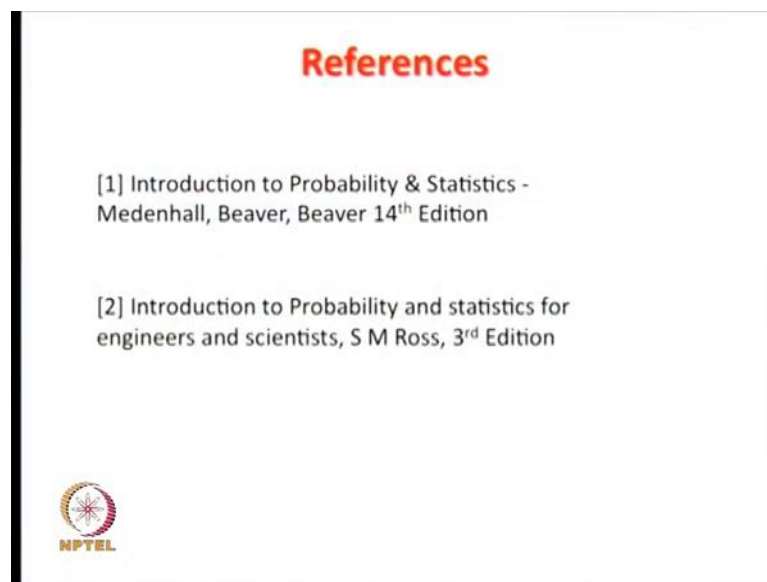


Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 01
Introduction to the course

Hi. Welcome you all to our course Introduction to Biostatistics. This is our first lecture, I just wanted to cover one lecture on what statistics is, what biostatistics is and then talk about few things which are important and which will serve as landmarks as we go ahead in this course. So, I will begin with the course material what is the courses that we are the two main reference books that will serve for you in this course.

(Refer Slide Time: 00:38)



So, this is the first book Introduction to Probability and Statistics by Medenhall, Beaver and Beaver and the second one Introduction to Probability and Statistics for engineers and scientists. So, both these books are available at on Flipkart, Amazon and on regular websites. So, I come to the first question, what is statistics?


(Refer Slide Time: 01:00)

What is Statistics?

The science that deals with the collection, classification, analysis, and interpretation of numerical facts or data

Examples

- Opinion Polls
- Weather Patterns
- Income as per professions

 NPTEL

So, statistics is the science that deals with the collection classification analysis and interpretation of numerical facts or data. For example, you know we continuously hear what the year selections right now and every now and then you see an opinion poll, what is the statistics by which Hillary Clinton will beat Donald Trump in the polls? You have the statistics which is repeatedly taken and you still come up with different companies which hold this polls which come up with different numbers. So, one person might predict that Hillary Clinton by Wind will mean by 30 percent words, another predicts by 20 percent words. So, this gives you the idea that this process is extremely complicated it is very easy to come up with a number saying this is the difference that is the difference, but there is a very active science behind this process in order to make sure that the numbers you pop out of the process are accurate.

Another example I want to talk about is the weather patterns every year in our country, we have the material ethical; you know the weather department which predicts whether the annual rainfall in our country will be as is it good enough for this year, will it beat expectations for the past one year, we have all known that in the has faced a severe drought for the last may be two consecutive years, this has been particularly extreme in parts of Maharashtra where I stay.

So, in all these cases, the prediction which these people; which these agencies give are really taken very seriously even for the government to make a informed decision as to

what should be the steps if the rainfall is adequate, what are the steps to make sure that the water is uniformly distributed; should there be a water cut in the case that rain is not up to expectations. Or if there is excessive rain should the dams release the water as a consequence of that there might be some you know flood like situation created in some other phase as has been recently observed in Bihar.

The last cases; for example, income as per professions; you know we continuously decide particularly when we are in finish our education in class 12, we ponder as to which in which you know trajectory we want to choose. Of course you know everyone wants to be an engineer the simple reason is we have enough industry positions open which after our educate after our bachelors we can get access to a job which will secure your future.

So, that is of course not true from every profession, but every profession has its challenges. So for example, if you want to be a singer then the trajectory you will go ahead in terms of income initially you will have very miniscule amount of income, but as you become established in your field you will have more and more income.

(Refer Slide Time: 03:55)


What is Biostatistics?

Application of statistics for understanding of biological systems or biological processes

the branch of statistics that deals with data relating to living organisms.

Examples

- Biology - evolution
- Medicine – drug resistance
- Public Health – Zika virus

 NPTEL

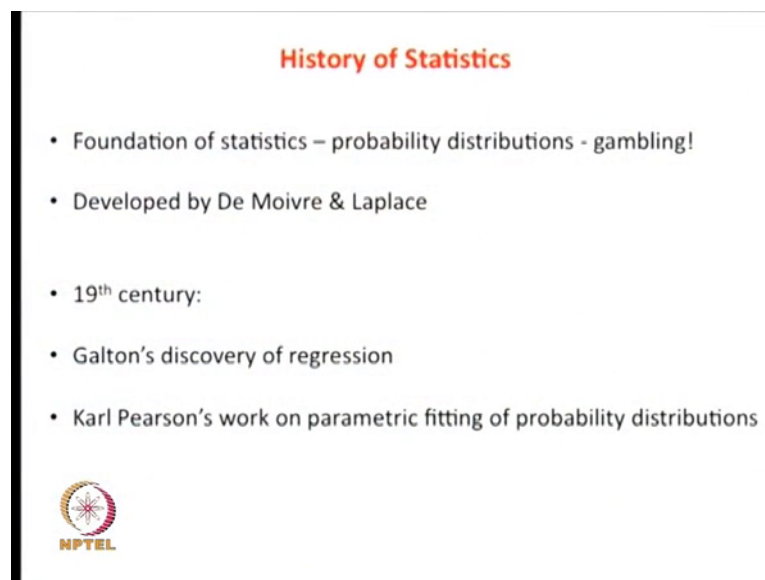
So, what is biostatistics? So, biostatistics is nothing but application of statistics for the study of living organisms for human beings for animals for any biological process for that matter. So, we can take examples from any field which concerns biology or medicine or bioengineering biology for example, we want to talk about evolution for

example, right in evolution we want to see by let us say comparing the structure of the bone how a dinosaur eventually became a bird so on and so forth.

And these kind of things you want to make exact analysis of specific structural components of a particular component of the body, let us say and see its evolution as a function of time in medicine right you want to predict the ability of drug resistance to arise in a given population given that in the past it has become resistant to this particular drugs what is the chance that they will become resistant to this new drug which is which is probably the company is thinking of bringing into the market.


Last case for example, is in public health you all been aware of the damage that Zika virus is currently producing. So, in the case of Zika virus, whenever you go to the airport and you will see that people coming from these spaces must have a mandatory health check up to eliminate the possibility that that person is a carrier of Zika virus. These are examples where statistics has been used to make some informed decisions first to measure, and based on that measure to make come up with some analysis, and then based on analysis to make some predictions as to what should be the corrective step.

(Refer Slide Time: 05:38)



History of Statistics

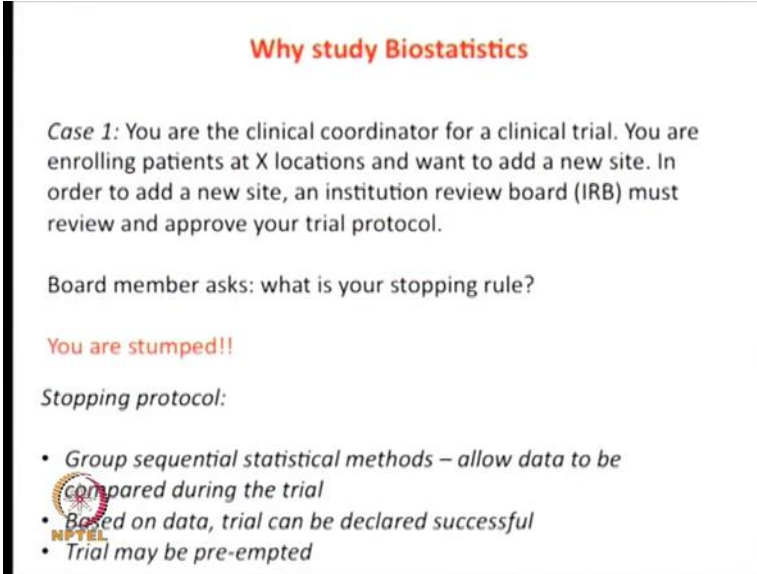
- Foundation of statistics – probability distributions - gambling!
- Developed by De Moivre & Laplace
- 19th century:
 - Galton's discovery of regression
 - Karl Pearson's work on parametric fitting of probability distributions

 NPTEL

So, if I look at the history of statistics it actually started with trying to understand the process of gambling which is one of the oldest professions people have participated in and this has roots in development of probability distributions which have been developed by De Moivre and Laplace in the 19th century. For example, you have Galton's

discovery of regression calculations work on parametric fitting of probability distributions and that has an affiliate you know statistics is particularly of particular relevance to the field of biology biomedicine and its applications in clinical science.

(Refer Slide Time: 06:15)



Why study Biostatistics

Case 1: You are the clinical coordinator for a clinical trial. You are enrolling patients at X locations and want to add a new site. In order to add a new site, an institution review board (IRB) must review and approve your trial protocol.

Board member asks: what is your stopping rule?

You are stumped!!

Stopping protocol:

- Group sequential statistical methods – allow data to be compared during the trial
- Based on data, trial can be declared successful
- Trial may be pre-empted

So, why study statistics let us take a sample case let us say your clinical coordinator for a clinical trial you are enrolling patients at X locations, let us say X is 5 or 6 as the case may be anyone you are pondering whether you want to add a new site for that for that clinical trial. So, for this for getting approvals for doing these clinical trials, you need to get approval from you know from the institutional review board. And one of the members of the IRB asks, you want to add one more location, but what is your stopping protocol if you have no idea, what is the stopping protocol?

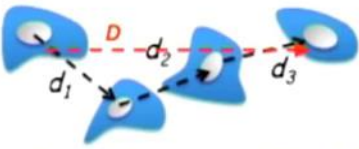
Well, stopping protocol is those set of rules which enable you to either declare your clinical trial a success at an earlier time or to priam the clinical trial if you already have some negative data coming because you collect data you know you know continuous manner. So, you have thresholds at various stages of data collection which enables you to label the clinical trial a success at an earlier time point or label it a failure. So, you want to minimize your losses and so on and so forth.

(Refer Slide Time: 07:31)

Why study Biostatistics

Case 2: Measuring cell motility – relevance to normal physiology (e.g., wound healing) & diseases (e.g., cancer)

Role of a specific protein?
Drug effect?



What to measure? – total distance travelled, net distance, instantaneous speed?
How to measure? – what should be the time difference between frames?

Another example, let us say of biostatistics is measuring cell motility. So, inside our body, cells are not sitting steady in our position, there many of the cells are continuously moving further take the case of you know take the case of immune cells they are continuously moving and scavenging for foreign particles which have come inside our body we want to the immune cells want to pre-empted. So, we want to the cells want to be able to move all over the place.

Now, let us say in the case of cancer or as the disease maybe you want to come up with a drug which inhibits this motion right cancer is a disease in which cells certain cells which are supposed to be in a certain organ they start migrating they start proliferating; that means, they had start dividing and then at some point of time they actually start invading the you know entire body they eventually reach secondary tissues and stop their function leading to death.

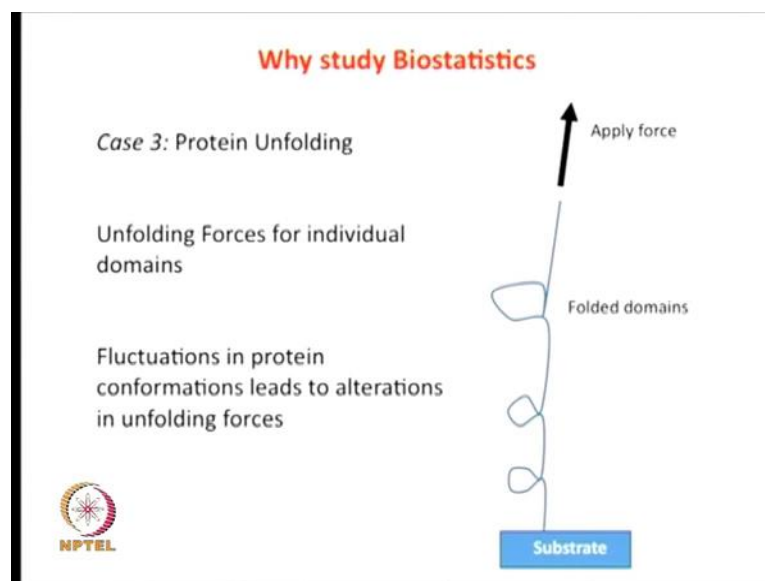
So, a significant amount of effort by scientists is in trying to find out drugs which target cell motility or the ability of the cell to move. So, you want to quantitatively assess whether this drug or drugs that you have currently developed in your lab whether they work whether they inhibit cell motility. And in your in your lab, you have this experimental assay in which you can have single cells which moves right which move randomly in your under the microscope you are acquiring these movies and the question here is right. So, you want to understand the role of a specific protein or the effect of the

drug as the case may be and so as a process you want to stop you want to measure the cell motility and see whether it has an effect or not an effect.

So, you want to essentially measure the cell movement. So, the question is how should we acquire the movie when the cells are moving let us say a cell is jiggling in one place, it is not moving. So, over the you know let us say within 5 seconds the cell is pretty much static in one position, but over 5 minutes it is actually moving by a distance equivalent to its own length. So, if you acquire every 5 seconds then essentially the information you are gathering is noise, but the information that you carry you can gather is if you are acquired after 5 minutes maybe that is an optimal time scale if you if you measure after 10 minutes that may not be good enough because cell might have come back to its original position and to you it seems that the cell is static.

So, how fast you acquire these images is examples of how you can you know statistics can help you in understanding what should be your acquisition sweet.

(Refer Slide Time: 10:20)



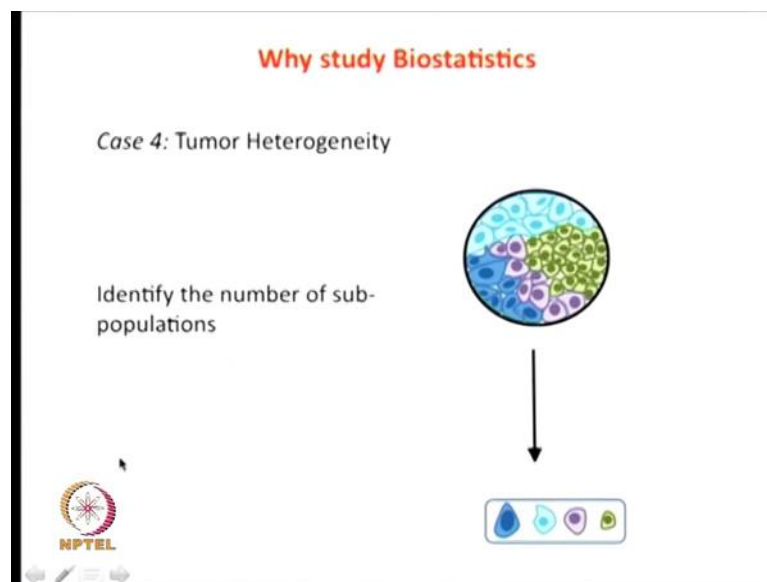
Let us take another example of a protein; proteins participate in various functions in our body when protein function is perturbed then they stop working and then they lead to disease processes now a protein you can imagine a protein like a long string which has loops in it. So, you have these loops, these loops are actually they hide some sites on them which other proteins ideally can recognize, but because these are in looped

configuration these sites are hidden and the cell cannot and other proteins cannot see which means they cannot bind and then the downstream signalling cascades are absent.

Under the conditions you want to understand what is the statistics of protein force, how much force do I need to exert for a particular domain of a protein to unfold for that what you can do is you can use techniques like atomic force microscopy to pull on the protein and what you will measure essentially is the force which is required to open up a given domain now the question here is. So, this now this force is not going to be a constant it will give you a distribution which means let us say for one particular experiment for a given domain of a protein you get a value of 10 Piconewtons in the next realization of the experiment you get a value of 12 Piconewtons.

The next one you might have value of 8 Piconewtons. So, it is. So, there are small differences in this, but if you take it enough number of times if you acquire enough number of data then what you probably will get is something called a Gaussian distribution or a normal distribution and the peak of the Gaussian distribution you can say is the unfold the force required to unfold that particular domain.

(Refer Slide Time: 12:09)



One last example of you know; an example of how biostatistics can be relevant. So, cancer again taking the case of cancer; cancer is consists of cells which are heterogeneous. So, as these different colours are used as to make out that you have cells of different types by different types it might be the cells are of different sizes the cells are


of the they are different extensive deformability one cell is stiff one cell is soft so on and so forth. So, you want to find out a process whereby you can understand and differentiate these phenotypic differences and then so statistic so you measure all these quantities for.

So, you have multiple parameters its size stiffness expression of some surface protein of interest so on and so forth. And then given for each cell type within the subpopulation you get its phenotypic characterization you want to understand what should be the combination of these metrics that would enable you to separate them into different lots.

(Refer Slide Time: 13:29)

Types of Studies

- Surveys & Cross-sectional Studies (reference point in time: now)
Texting Patterns in College Campuses
- Retrospective Studies (reference point: past)
Trying to figure out what caused a suspected outbreak of food-borne illnesses
- Prospective Studies (follow subjects from present to future)
tracking chronic diseases with long latency periods



So, again here the role of statistics will help you in identifying the number of sub populations and their properties. So, what are the types of studies which fall under statistics what kind of you know how do how do statistician acquired data for example, what the most common is surveys and cross sectional studies. So, the important point about surveys is they want to collect information pertaining to the population which is right now, right here, right now.

So, it for example, a particular company let us say Airtel might want to know what is the patterns of texting in college campuses by students right how many number of text messages you know do students on an average send. So, that they can accordingly decide you know how many free SMSs can I Airtel offer versus after how many times to the charge. So, if the overcharge then essentially, the student population will go ahead and sign up for some other company so you want to retain your; you know retain your

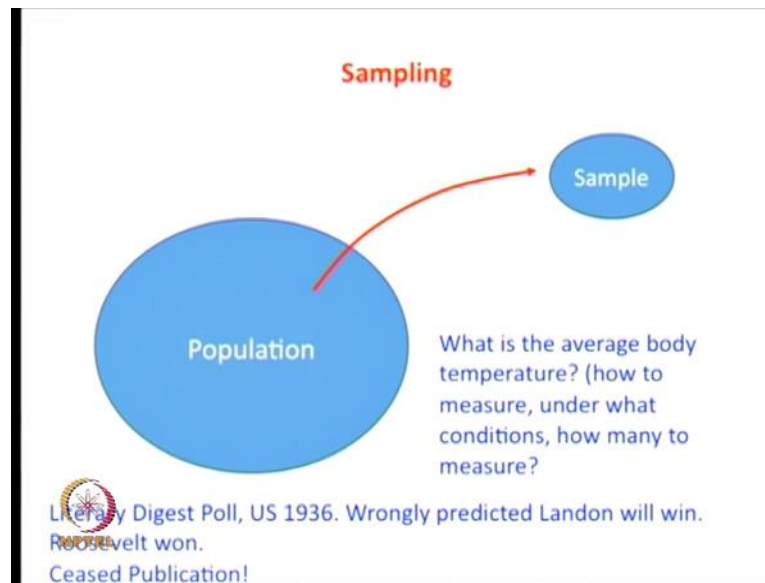
consumer base yet maximize your profits. So, survey is a critical example of how this might help.

Another case is a retrospective study. So, retrospective study means you have a population of people who might have been exposed. So, something let us say yesterday I went to a restaurant particular restaurant, I had my food and today there is a food allergy or an outbreak of you know food poisoning. So, there is this investigative agency they want to know whether is it, because of one particular food that we have consumed because of which this food wise poisoning is occurring or is it a random event in that case they would want to peep you want to sample people who have visited that particular restraint on that given day and compared with other people of similar demographic background, but who have not gone to that restaurant. So, this is an example of a retrospective study.

Another thing can be a prospective study saying prospective studies these are studies associated to see for example, in college campuses as you know the when the student population starts smoking and you want to do at what point of or how long of smoke smokers will eventually lead to a chronic disease like a lung cancer or anything, but which have long latency period. So, that disease does not manifest itself in 10 days or 20 days what we are talking about years maybe 10 years or 15 years. So, steady stations preferably like to track certain people over time to see how they are you know what is their incidence of lung cancer or what is their current state as a when they start is smoking to that current state and of course, the final sample you know very often used particularly for biostatistician is clinical trials.

So, clinical trials are performed. So, after you have proved that a drug works in the lab before humans we humans start taking that medicine it has to be proven two things one that if I consume that particular medicine I am not going to die. So, I do not suffer any extreme diseases or side effects and to the medicine actually works. So, that it can be put to the market.

(Refer Slide Time: 16:52)



So, there are lots of studiously how you know what should be the; my study size in order to test whether a particular sample is a drug is working or not working so on and so forth. So, in all these processes there is a critical component of sampling in essential you have a population which is this big. So, in India for example, we have huge population and let us say I want to make a soap or Patanjali wants to come up with its new toothpaste or some other product it wants to test on a sample right it want to do that measurement on that sample and ask to test whether my product is good or bad what should be the price of the product so on and so forth; I have to have that questionnaire in terms of a survey, but I want to ask some people.

Now, how many people should I ask is it 10, 20, 200, 2,000, so of course, you cannot ask every or dumb person on this earth or every person on in a city you are still going to sample again what is your sample you know the kind of population from the population you might draw multiple samples for cities like Mumbai, Delhi, Madras, you might to draw certain kind of people for to get feedback on how that product might do in a more town like area you might go too much smaller towns and the statistics the information you will get might be completely different, but say as an exception let us say if you are just doing it on college campuses maybe it may not be that different between a city or a town. So, choosing your sample and sample the type of sample and the sample size represent two of the most important choices that we must take.

So, and as an example of how you know wrong choices can lead to complete mess. So, there is literally there is a magazine in the u s colour literally digest. So, it conducted a poll in 1936 and our; the presidential poll and wanted to ask who would win Landon or Roosevelt. So, the particular magazine came up with the prediction then Landon will have will be victorious, but which did not happen; not it not only was it a borderline case, but it was a significant victory for Roosevelt.

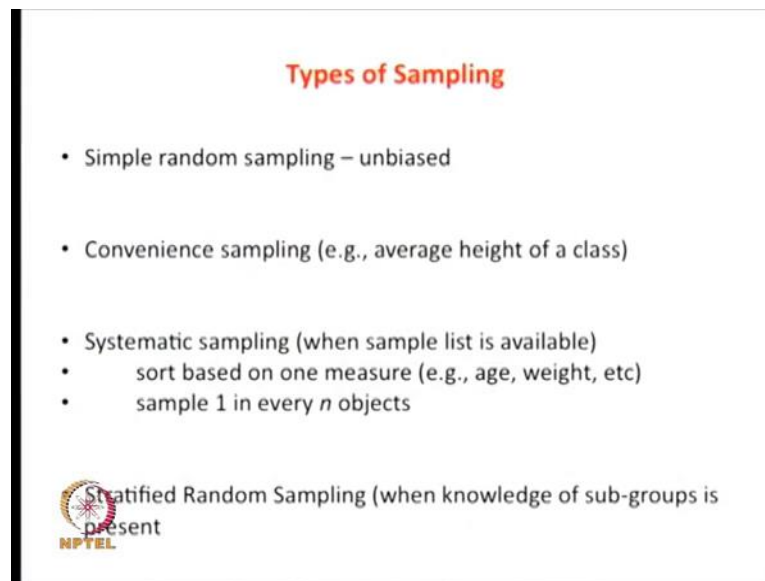
So, this was a clear failure for that particular pole and no wonder the publication actually stopped shortly. So, it is very important to make predictions which are reasonably accurate and we see repeatedly in India in election cycles you have every TV every major channel has its own prediction as to what it happen and still in spite of you know all the checks and balances still there is reasonable amount of heterogeneity in the prediction as to who will win and by what margin they will win?

So, that brings us to sampling and what is the type of sampling. So, in terms of type of sampling; the most simple thing is simple random sampling which means that I want an unbiased view of the population which is. So, let us say in classes, I want to measure the average height, I average height of the class, I just ask 5 random students, I you know I maybe I even blindfold myself, I randomly walk around because you know speaker student and measure his her height that will give me an average height.

Now, imagine I am doing the exact same thing, but I only choose only girls or only boys. So, of course, my measurement is going to be skewed right because on an average boys are slightly taller than girls. So, my average height number that I come up with is going to be erroneous the other sample is systemic sampling. So, let us say you already know the population or the sample list for example, let us say the class is the population and I want to measure I want to measure 10 out of 10 measurements whatever be it height weight or you know some other age from the population and I know population size is 50. So, what I do I arranged the entire population by name by let us say our it is ascending or descending order the surname and out of the 50, I choose number one then I want you know 10s roughly your spacing is 5, I chose number 6 so on and so forth. So, there is randomness and still I am not biasing myself and, but I have kind of made it in a much more systematic manner. So, if your population size is reasonably known then this systemic sampling is very commonly used.


The last one is stratified random sampling let us say for example, you have a basket of balls which are of three types of three different sizes right and on an average you know what is the ratio of the total number of red balls to blue balls to yellow balls right three colours let us say balls of three colours and you know the size the stoichiometry of red to blue to yellow then logic would dictate if you are doing a random sampling then the number of balls you would draw would be in proportion to them proportion to the colours. So, this is called stratified sampling.

(Refer Slide Time: 21:56)



Types of Sampling

- Simple random sampling – unbiased
- Convenience sampling (e.g., average height of a class)
- Systematic sampling (when sample list is available)
 - sort based on one measure (e.g., age, weight, etc)
 - sample 1 in every n objects
- Stratified Random Sampling (when knowledge of sub-groups is present)

 NPTEL

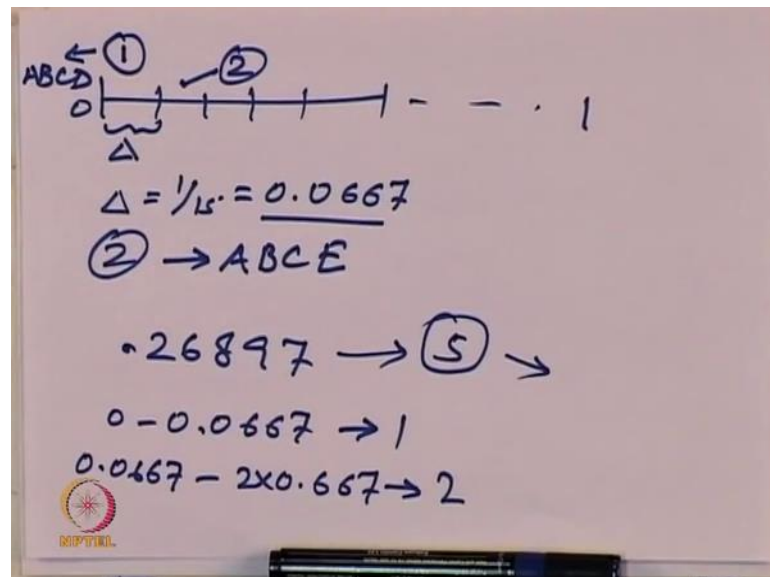
Let us get over a very simple you know do a very simple case of simple random sampling of. So, let us say you have 6 patients and I want. So, and let us say we name the patients A, B, C, D, E, F and we want to pick 4 of the 6 patients. I want to pick 4 of 6 patients. So, what do I do what I can do is I you know and this I want this process to be completely random. So, what I do is I choose. So, I and. So, if you have a b c d e f they can be total of 64 or 15 combinations possible in which you have 4 different patients being selected. So, let us say A, B, C, D is one example, A, B, C, E is one example, A, B, C, F is one example. So, if you write down the details you will see that there are roughly 15 combinations possible.

So, I want to draw 4 patients from 15 samples and I want to ensure that it is completely random what I do I assign A, B, C, D. So, let there are 15 of these combinations possible to each combination I assign a unique identifier let us say. So, A, B, C, D is identified,

number one A, B, C, E is identified number two so on and so forth. Now from these, so what I do? So, I have total of 15 numbers now imagine I break down the number one; 0 to one that range into 15 equal intervals. So, your interval your; you know the size of each interval is going to be 0.067.

So, the size of each interval is going to be 0.067. So, what you do. So, you take up a uniform 5; you know uniform random number table, you choose any particular number and convert it into decimal let us say we chose a particular number it is 2 5 6 9 7. So, I can just put a decimal with the front that is 0.25697.

(Refer Slide Time: 24:14)



Now, given that I have broken the entire range. So, if I have broken the entire range 0 to one. So, I have broken the entire range 0 to 1; 0 to 1, I have broken down into delta. So, this delta is equal to one by 15 equal to 0.0667. So, whenever I choose when I am this I choose my identifier one which is the number which is the combination A, B, C, D when I choose 2, my two is my combination A, B, C, E so on and so forth. So, from the table I had chosen a number let us say 2 6 8 9 7, I converted into decimal place.

So, I know this. So, decimal place means my delta is 0.0667. So, 0 to 0.0667, I choose identified 0.0667 to twice of 0.0667 I choose identified 2. So, like this, I think this 2.26897 should be identified 5. Accordingly, I can choose one particular combination of the patient. So, this is a way of doing simple random sampling.

(Refer Slide Time: 25:32)

Simple Random Sampling

- Choosing 4 patients from a group of six patients (A, B, C, D, E, F)
- Number of possible ways: ${}^6C_4 = 15$ (enumerate)
- Take interval from $[0, 1]$ and divide into 15 equal parts ($= 0.0667$)
- Look up table

Col./Row	1	2	3	4	5	6	7	8	9	10
1	00439	60176	48503	14559	18274	45809	09748	19716	15081	84704
2	29676	37909	95673	66757	04164	94000	19939	55374	26109	58722
3	69386	71708	88608	67251	22512	00169	02887	84072	91832	97489
4	68381	61725	49122	75836	15368	52551	58711	43014	95376	57402
5	69158	38683	41374	17028	09304	10834	10332	07534	79067	27126
6	00858	04152	17833	41105	46569	90109	32335	65895	64362	01431
7	86972	51707	58242	16035	94887	83510	53124	85750	98015	00038
8	30606	45225	30161	07973	03034	82983	61369	65913	65478	62319
9	93864	49044	57169	43125	11703	87009	06219	28040	10050	05974
10	61937	90217	36708	35351	60820	90729	28489	88186	74006	18320
11	94551	69538	52924	08530	79302	34981	60530	96317	29918	16918
12	79385	49498	48569	57888	70564	17660	68930	39693	87372	09600
13	86232	01398	50258	22868	71052	10127	48729	67613	59400	65886
14	04912	01051	33687	03296	17112	23843	16796	22332	91570	47197
15	15455	88237	91026	36454	18765	97891	11022	98774	00321	10386
16	88430	09861	45098	66176	59598	98527	11059	31626	10798	50313

Choose one at a TIME

So, that is you know pretty much what would give you an idea this is just the background idea of what you know means statistics and so to summarize, we started off the definition of statistics which is a signs statistics is nothing, but the you no signs of numbers and biostatistics is application of statistics to the field of biology or you know biological processes or biological systems and it has you know applications in basic biology in medicine in healthcare in public health so on and so forth.

So, after that we discussed about you know some examples of how statistics might be beneficial in our; you know understanding of biological systems. So, I give you an example of clinical trial how to design a clinical trial and what to understand from that I give you an example. So, it you need to understand how clinical trials are designed right what is the sample size of a clinical trial so on. And also be acquainted with the jargon of clinical trial I give you the example of you know cell motility you know this is an example where statistics help you to understand that at what rate should you acquire movies. So, that your movies are exactly you know. So, the movies are not artefact or not noise do not give you noise when something is just jiggling in place versus something is data; that means, cells actually moving from one place to another, but you do not also do not want to miss you know too much of the information. So, if you sample to infrequently then you are losing information as well.

So, statistics will help you to get that information we discuss the case of protein folding protein unfolding and how force can be used to open it and you can measure what is the average stability of a domain by doing experiments pulling the data generating a distribution and getting to understand what this means and last talked about you know how you collect these samples. So, in terms of sampling techniques random sampling or bias sampling and you want to do it retrospectively as a survey so on and so forth.

So, this is just our first lecture. So, know from our next lecture we will start talking about how to measure numbers you know from an experiment, from the raw data, how do you go about generating a plot that is very important. And how to represent that plot so that the common man or you know someone in the audience with scientific non scientific they can understand, how to you know, what is the gist of the data. You cannot just show the entire raw data, but the useful matrix from that raw data. So, first step of that is actually to plot and convey what it is.

With that, I thank you from the first lecture. I will also upload us you know few MCQs so you can get an idea of what we discussed and get to refresh in your memories of this class.

Thank you.