

Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 03
Data types/Binomial Distribution

We will continue on the course on Biostatistics and Design of Experiments. As I mentioned in this course, I am going to talk about biostatistics; that is the part one of the whole thing and then comes the design of experiments.

(Refer Slide Time: 00:17)



Biostatistics
Introduction to statistics (various distributions)
Normal distribution
t distribution
Z distribution
Binomial distribution
Poisson distribution
Weibull distribution
Confidence interval
Test for normality
Tests of significance - t test (one sample, two sample)
F test
ANOVA (one way, two way..)
χ^2 test/Odds ratio
Non parametric tests
Other tests

In Biostatistics, we are going to look at large number of distributions like Binomial distribution, Poisson distribution, Weibull distribution, **T**-distribution, Z-distribution, Normal distribution and so on. Then we are going to look at something called Confidence interval, Test for normality, Tests of significance, different types test, t-test, F test, ANOVA test. And under t test you have one sample t- test, two sample t-test and then we are also going to look at Chi square test or Chi square distribution. Then we are going to look at Non parametric tests, other type of tests that are possible in biostatistics which does not need to have a Normal distribution.

(Refer Slide Time: 01:07)

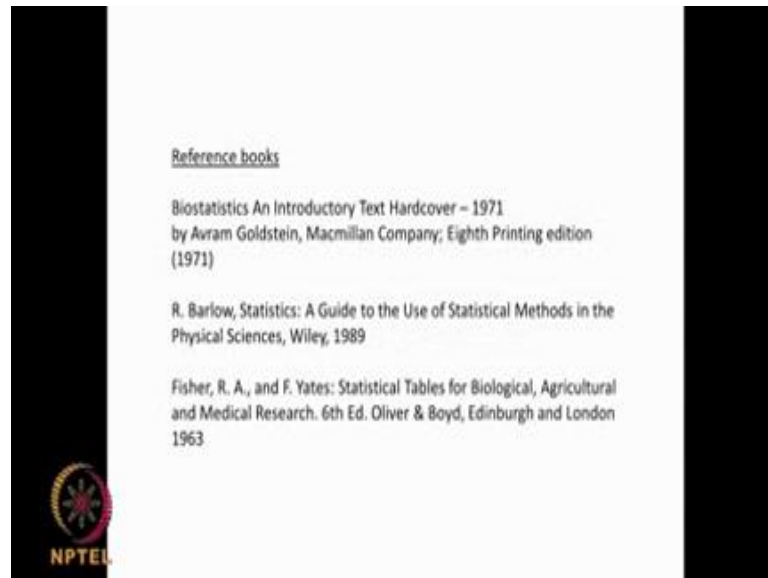
Design of experiments
One factor at a time
Full factorial design
Fractional factorial design
Confounding/alias
Screening design
Second order design
Regression analysis/mathematical modelling
Data reduction



Then under design of experiments, we are going to look at one factor at a time. How do I change one factor, like if I am changing temperature alone, then after I finish temperature optimization, I go to **pH** alone, that is called one factor at a time, and then go into a design, Full factorial design, where I am changing many factors simultaneously, then there is something called Fractional factorial design, where you are doing a fraction of the full factorial design that means you are cutting down on the experiments.

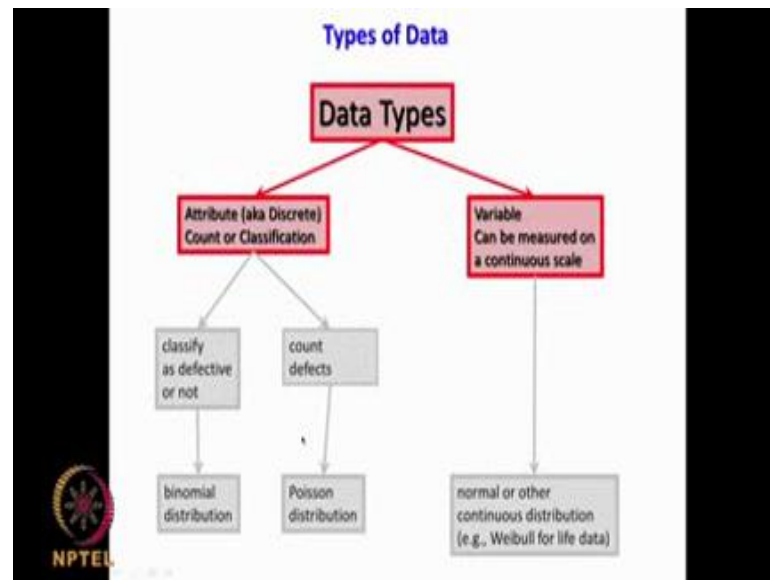
Then you are going to talk about what is this Confounding and alias and how does confounding affect when you start doing the design of experiments. Then there is something called Screening designs; that is initially you start looking at a large number of parameters and carry out experiments, that is called Screening designs. Then you come to Second order designs, that means non-linear type of designs. Then once you collect the data from the design you do a Regression analysis, mathematical modeling. Then finally you go into data reduction that is the second part of the course. The first part is Biostatistics second part is Design of Experiments.

(Refer Slide Time: 02:13)



Let us get into and before that these are some of the books which may be very useful for you; Biostatistics An Introductory Text by Goldstein, then we have Barlow, A Guide to the Use of Statistical Methods. Then Fisher and Yates, this is very very good book which gives lot of statistical tables. Because as you go along you will come across lot of tables t tables, f tables, random numbers, odds ratios, confidence intervals, p values and so on, for all these you need to have some tables and this one gives you. Of course, you can get the tables online also but they are all based on this particular book called Fisher and Yates. They have developed statistical tables for Biological, Agricultural and Medical Research.

(Refer Slide Time: 03:02)



Let us look at Data Types. There are 2 types of data, one is called the Attribute data other one is called the Variable; continuous data. Attribute data could be like 0-1, pass-fail, live-dead, black-white. It is like a numerical numbers, it could be in counted, 10 defects in 10,000 samples, 10 failures in a class of 100 students, you can count it, you can classify it. So, it is based on numerical numbers, it is discrete. Whereas in Variable data; continuous data, you can have continuously changing, for example, I can measure temperature of a fermenter continuously, I can call it 26.5 or 26.6, 26.7, 26.8 like that I can measure the temperature very continuously. Similarly, I can measure the **pH** of the solution in a very continuous manner 3.1, 3.2, 3.3 and so on, that is called the continuous data. So, we have the discrete data we have the continuous data. So, any data type can be divided into these two forms.

ssssss

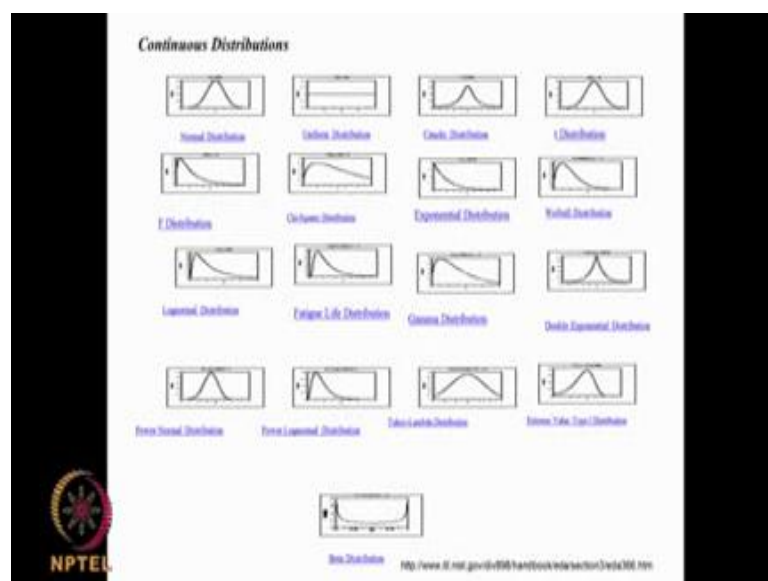
Now under this discrete, we can classify it as defective or not. Especially, if you take a factory where they manufacture lot of product. For example, they are manufacturing screws and they would like to have the screw of 10 mm diameter, so any screw that is not 10 mm; if it is 9 mm or if it is 11 mm it is called a defect. We can say out of 1 million screws that are manufactured in each week there could be so many screws, 10 screws which are defective, that means they do not have 10 mm as the diameter, the diameter

could be different, that is classifying, so you have something called the binomial distribution coming in to picture. So there are 10 defective screws out of 1 million screws that are manufactured in this particular week.

Then we also have something called Poisson distribution, this is again giving a count. There are 3 road accidents in the city of Chennai in a months' time, there are 4 people suffering from HIV in this particular village in South India, you are giving some numbers. And again, the numbers are collected based on large number of samples. Again, it is count that is called Poisson distribution. We have the Binomial distribution, out of 10,000 samples 10 are defective or we have the Poisson distribution, where I am saying there are 3 deaths per day in the city of Chennai.

Then under the continuous data we have the Normal distribution. You must all heard of normal or the uniform distribution which looks like a bell type of curve and also we have the Weibull distribution which discusses the life of, say for example, a light bulb or a fan or a refrigerator that is called the Weibull distribution. So, that is continuous data, we can measure the data continuously that is called the Weibull distribution. So, we have two types of data, the discrete data and the continuous data. Discrete data is used for identifying how many defects are there in a sample and how many accidents are happening in Chennai per month or per week and so on, whereas continuous data we are measuring the data in a very continuous manner.

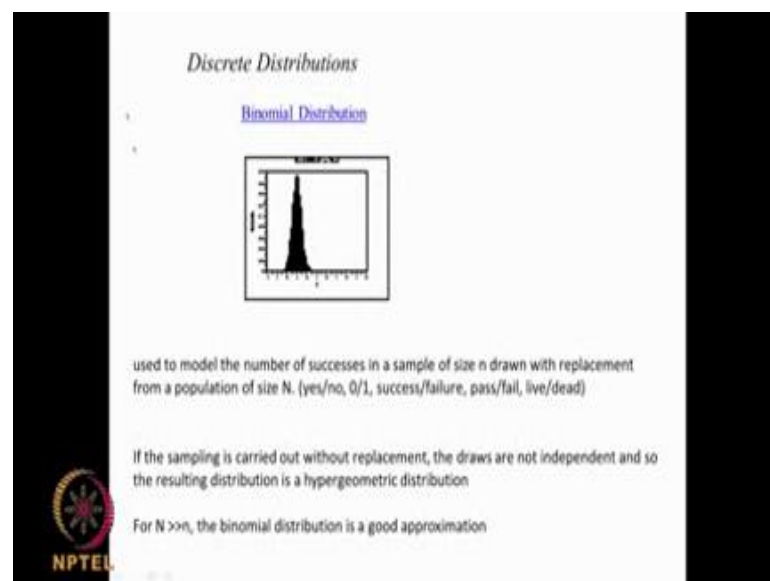
(Refer Slide Time: 06:36)



In fact there are large numbers of distributions much larger than what I talked about. As you can see, do not get scared we have Normal distribution, we have Uniform distribution then we have t distribution we are going to talk about this. We have F-distribution we are going to spend some time on this, then Chi square distribution we are going to spend some time on this, Weibull distribution and so on actually. As you can see these all are continuous distribution, fatigue life distribution, gamma distribution, double exponential, power normal, power logarithm, beta distribution so on. So, large numbers of distributions are there. They are used in different scenarios, different requirements, different problems, but I will be spending time on normal, I will be spending time on T-distribution, I will be talking about the Chi square distribution, F distribution, Weibull distribution but all these distributions are very useful actually.

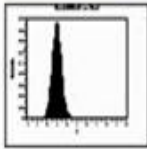
As you can see they have different shapes and that means the probability of certain event happening will follow different type of relationship or mathematical formula. So, these all are continuous distribution. In the discrete we have the binomial and we have another type of Poisson distribution, so these are continuous distribution. I said we will spend time only on few of these not all of them.

(Refer Slide Time: 08:05)



Discrete Distributions

Binomial Distribution



used to model the number of successes in a sample of size n drawn with replacement from a population of size N . (yes/no, 0/1, success/failure, pass/fail, live/dead)

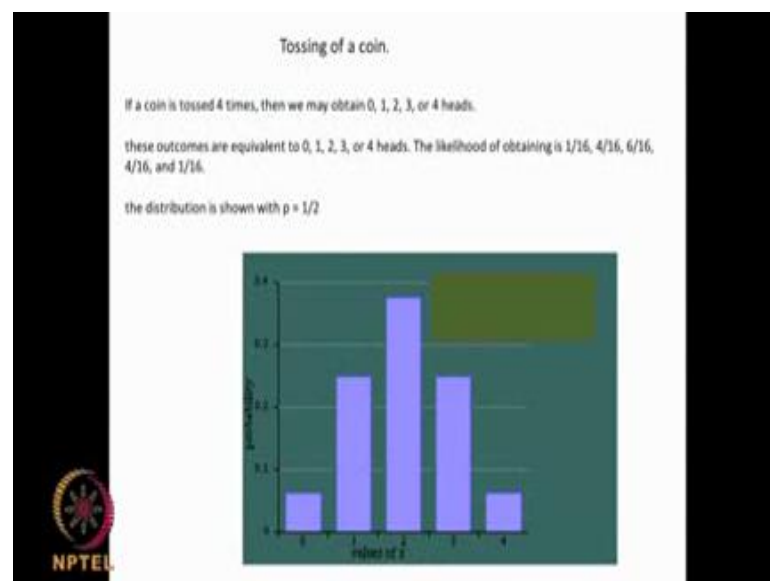
If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution

For $N \gg n$, the binomial distribution is a good approximation

NPTEL

Let us check Binomial Distribution. You must have all read in your school talking about probability, tossing a coin, tossing a dice and so on actually. Binomial Distribution is based on yes-no, 0-1, success-failure, pass-fail, live-dead, black-white or suppose I have a dice which has 6 faces then I am throwing a dice you may get a number 1 or 2 or 3 or 4 or 5 or 6 at equal probability, all these are based on Binomial Distribution. The sampling is carried out without replacement that means you are not putting it back, the draws are not independent. So, binomial distribution is a good approximation here. If I am tossing a coin probability of coin showing head could be half, probability of coin showing tail it could be half. If I throw it 10 times same coin, if I want to know what is the probability, 4 of them out of this 10 is heads I can use the binomial distribution or if I am going to say that the birth defect of children born in India is 10 percent and I go to a village which has got 1000 children what is the probability that 4 of these children will have that particular defect, both defect then I can use the Binomial Distribution. So, that way Binomial Distribution becomes very useful for us to do.

(Refer Slide Time: 09:34)



Let us look at a simple problem, Tossing of a coin. I have a coin as you know I can get either heads or tails. That means equal probability, 50 percent probability for heads 50 percent probability for tails. So, I toss the coin 4 times, I can get 0 heads that means all of them become tail, I can get 1 head that means I can get 1 head and 3 tails, I can get 2

that means 2 heads and 2 tails, I can get 3 heads and 1 tail or 4 heads and no tail. All these are possible and the likelihood of getting each one of them is given by this formula 1 by 16, 4 by 16, 6 by 16, 4 by 16, 1 by 16. How do we get this?

In the next slide, I will show you the formula. This is how the distribution will look like I toss the coin 4 times obviously getting 2 heads, 2 tails is most probable. How to get this number of 6 by 16, I will tell you in the next slide. And then getting 1 head and 3 tails or getting 3 heads and 1 tail are equally probable which comes second and then getting 0 heads or getting 4 heads again is less probable in this but they are equal. This is how the binomial distribution will look like. Now what is the formula for calculating this probability let us show it in the next slide.

(Refer Slide Time: 10:59)


The probability of getting exactly k successes in n trials is given by the [probability mass function](#):

$$f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

If we want to determine the probability of getting 0 heads then
 $n = 4, k = 0, p = \frac{1}{2}$...

$$f(0) = \frac{4! (1/2)^0 (1-1/2)^{4-0}}{0! (4-0)!} = 1/16$$

If we want to determine the probability of getting 2 heads then
 $n = 4, k = 2, p = \frac{1}{2}$...

$$f(2) = \frac{4! (1/2)^2 (1-1/2)^{4-2}}{2! (4-2)!} = 6/16$$


This is how the probability equation looks like. The probability function $f(k)$ is given by

$$f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

So, n trials, k successes, p is the probability.

So, n times you are doing something and k is the successes you are talking about, p is the probability. In the previous problem like I am tossing the coin 4 times, n will be equal to 4. If I want to know what is the probability for 0 heads, then k will become 0 and p is half because I can get either head or tail. Probabilities p which is half n will be 4 and if I want to get a zero heads what is the probability I want to calculate; I will put k as 0.

n factorial you all know must have studied, n factorial is nothing

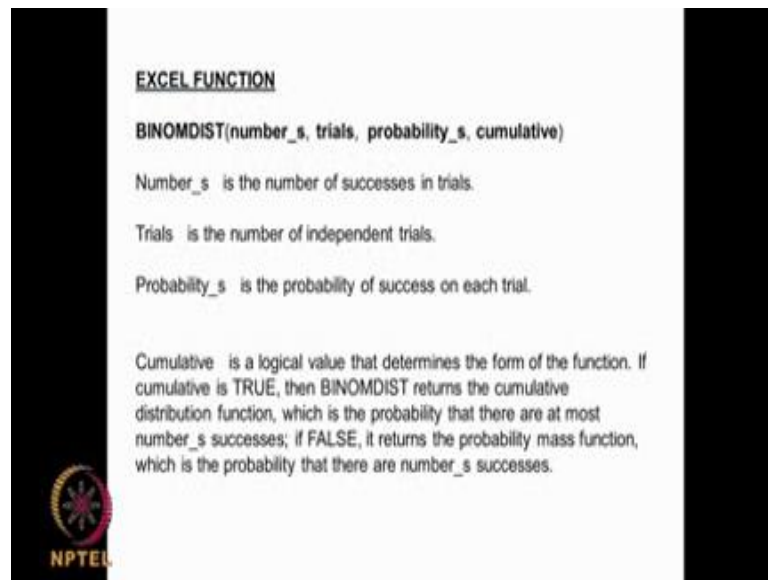
but $n(n-1)(n-2) \dots 3$ and so on. When I put $k=0$, I substitute here, I will get $4!$ and the denominator I put $0!$ then I put $4-0!$ $(1/2)^0$, $(1-1/2)^{4-0}$, that is what I have written here. $0!$ is 1, $4!$ is $4 \times 3 \times 2$ that is 24, $(1/2)^0$ is 1, $(1-1/2)^4$ is $1/2$ of raise to the power 4, this is $4!$ at the denominator. So, these two will cancel, these two will cancel. So, we have $1/2^4$ that means $2 \times 2 \times 2 \times 2$, 4 times that is $1/16$. If you want to see 0 heads when you toss a coin 4 times the probability will be $1/16$. You see that is what I had mentioned here right, $1/16$. This is how you get the data.

Now if you want to know what is the probability to get 2 heads when I toss the coin 4 times, so $n=4$, $k=2$ and again p will be half, you put $4!$, $2! 4-2$ $4! (1/2)^2$, $(1-1/2)^{4-2}$

. So you do all these calculations, you end up with $6/16$, I mentioned here $6/16$ that is the maximum. When you toss the coin 4 times what is the probability of getting 2 times head in that 4 is $6/16$ that is the maximum.

Like that if you want to know with 4 times tossing if $k=1$ that means 1 head, what is the probability of getting 1 head when I toss the coin the 4 times. I put $n=4$ but I put $k=1$, p will be $1/2$ in all these cases, $0!$ you should remember is always 1. It is simple to calculate.

(Refer Slide Time: 14:17)



EXCEL FUNCTION


BINOMDIST(number_s, trials, probability_s, cumulative)

Number_s is the number of successes in trials.

Trials is the number of independent trials.

Probability_s is the probability of success on each trial.

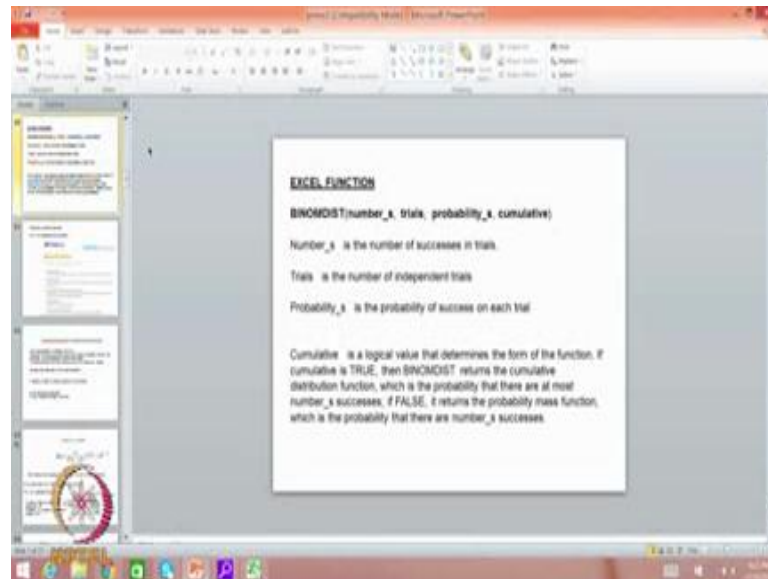
Cumulative is a logical value that determines the form of the function. If cumulative is TRUE, then BINOMDIST returns the cumulative distribution function, which is the probability that there are at most number_s successes; if FALSE, it returns the probability mass function, which is the probability that there are number_s successes.



Now you can do the same calculation using Excel as well. Excel has a function called Binom Distribution, there are 3, 4 terms inside this. Number s is the number of successes in the trials, trials is the total number and probability s is the probability and cumulative you can say true or false. If it is false it will give you the exact number whereas if you put true it gives you the cumulative number. Trials is n, in the equation, number s is the success k, probability is your p small p and here we put true or false, if we put false it gives you the exact answer. For example, in the previous problem where we looked at 4 times I tossed the coin, I want to know 2 successes with heads, what will I do, I will put 2 here, I will put 4 here, I will put half here and I can put false here and that will give you the Binomial Distribution answer, I should get $\frac{6}{16}$ as my answer. Let us look at it in the Excel as well.

(Refer Slide Time: 15:44)

SSSSS



This is the function I said it is called Binom Distribution, Number s is the number successes, I will put 2 here, number of times I do that is 4 here, probability is half that means I put 0.5 here and if I put false it will give you the exact answer whereas when I put true it will give you summation of all the answer. I will put false, what did I get? I got 0.375, now is it same as 6 / 16? See! That is 0.375. Using excel we can calculate the binomial distribution as you can see this is the equation, for example, this is the number of successes, this is the number of trials this is n, this is k, this is the p and here we put false to get the exact answer here. When you put true it adds up it is a cumulative answer that means, if I put true here it tells you what is the cumulative probability for getting at least 2 heads out of 4 trials that means it will look at 0, it will look at 1, then it look at 2. It will give you the summation of all these three things. We can use excel also to do the same calculation or we can actually calculate it out also. You understood. You have a excel function called Binom Distribution and there are 4 terms here, the trials this is equal to n, this is equal to k, this is equal to p and here you put false to get the answer. Now there are many softwares which can do this job also, some of them are commercial, there are could be something free also in the net and so on.

(Refer Slide Time: 17:46)



I also looked at software and there this free online statistical calculator and this is the link for that you know GraphPad, it is called GraphPad software. It can do lot of nice calculations online, we put in some data and it can do some calculations. I am going to use this and we are going to, we can do some of the problems using this. This online software as we can see can do lot of calculations it can look at Binomial Distribution, Poisson Distribution, Normal Distribution, it can look at different types statistical test, t test, f test and so on, we are also going to use this. This is the link to that [www graphpad dot com quickcalcs](http://www.graphpad.com/quickcalcs).

Let me show you that here, when you do that as you can see here, this is the GraphPad QuickCalcs. We have the Binomial Distribution coming into picture, we click on it and then we go continue, when you continue as you can see here calculate different types of distribution. Let us go into binomial, we will talk about different distribution later as I said I am going to talk about Binomial, I am going to talk about Poisson, t distribution, normal and so on. Here you have the Binomial, so we say continue. Here we have the Binomial Distribution. How many trials? We are doing 4 trials. What is the probability of success in each trial 0.5, calculate probabilities? Here you can see it gives you everything. So, number of successes 0 means it gives you 0.25 that is 0 heads out of 4 trials the probabilities 6.25 and then if you are talking about 1 success out of 1 head out of 4 trials, you get a 25 % but here you gives you the cumulative, 6 + 25 is giving 31. How do you, even in Excel if we put true as the last term you will get the cumulative,

whereas if you put false you will get the exact. 2 trials you get 37 percent, you can see 0.375 and the cumulative will be some 6 + 25 + 37, 68% So, 3 successes out of 4 it gives you 25 %, cumulative wise it is 93 %. All 4 heads out of 4 trials 100 % it gives you. We can use this particular online software also there could be many online softwares but I am looking at this particular online software because it looks good. There are many commercial softwares also one can go about using them, it depends upon whether you have the availability of these. There are softwares which may be even freely downloadable but this is simple online software where you give the data and it gives you the results. As you can see here in our problem we had tossing the coin 4 times and you can get heads or tails with the probability of $1/2$ now out of this 4 times, 0 heads 6.25 % probability. Out of 4 times 1 head 25 % probability. Out of 4 times 2 heads 37.5 % probability. Out of 4 heads 3 heads 25 % probability. Out of 4 trials, 4 heads 6.25.

I showed you 3 different ways by which we can calculate this. Right. One is using this equation if the data is very small we can use this and do it that means if the k , n and all is small we can. Otherwise we can use the excel function which is called Binom distribution where this is the number of successes that is k , number of trials that is n and this is the probability that is in this case $1/2$ p then here we give false or we can use this free online statistical calculator which I showed you. We click here and then we give number of trials as 4 and probability is $1/2$, it gives you the entire table for 0 success out of 4. What is the probability for 1 success out of 4? What is the probability for 2 success out of 4? What is the probability and for 3 success out of 4? What is the probability and that is what it is giving you here, right? As you can see it gives you in the entire table. So, I showed you three different approaches by which we can do the Binomial Distribution calculation.


(Refer Slide Time: 22:33)

A Biological Application of the Binomial Distribution

1% of the population is infected with HIV+.
There are no obvious symptoms that can be used to recognise carriers, thus individuals must be selected at random and tested.
If the sample size is too small there is a risk of not finding any carriers,
too large then resources will be used inefficiently.

A decision is made to obtain a sample of 20 individuals.

Is this sample size adequate?
Will any infected individuals be found?




Let us go further, let us look at a biological application. 1 % of the population is infected with HIV plus I am just giving. So, may be in a country 1 % of the population is infected, there are no obvious symptoms that can be used to recognize the carriers. We assume that if I look at somebody I cannot tell whether the person has HIV or not unless I do a detailed study. For example, I need to select some people and do a detailed study if the sample size is too small then I might not be able to find at all then if I take a very big sample then I need to do lot of sample collection sample analysis. I need to spend lot of money that is also inefficient. So, what do I do? Is it ok if I just take 20 people? Is this sample adequate? Will I be able to find at least 1 percent in that? That is problem. How do I do using Binomial Distribution?

(Refer Slide Time: 23:33)

$$n = 20, k = 0, p = 0.01$$
$$f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$
$$= 0.82$$

82% chance that a sample of 20 individuals will fail to find any infections



I can use $n = 20$, I can say $k = 0$ that means I am in that 20, I am not finding anybody with that and $p = 0.01$ because I said 1 % of the population. So, $p = 0.01$. When I put it in Binomial Distribution, 20 factorial because $k = 0$, this two will get canceled out, $p = 0.1$, $k = 0$, this also will get canceled out. So $1 - p$ is 0.99^{20} gives me 0.82. What does that mean? There is 82 % chance that if I take 20 people, I will not even find 1 person with that disease. Did you notice that? It is very very important finding, there is a 1 % population is infected with HIV but if I take 20 people randomly, there are 82 % chance that I will not find anybody with that in that sample of 20. So, I may say nobody is infected, obviously what does it mean? My sample size is too small or if I can say $n = 20$, $k = 1$ then I can do the same study and see what is the probability of finding at least 1, what it means is when I randomly select 20 people, I am not able I will not be able to find even 1 % with that particular disease. So I may say that nobody is infected with this particular disease.


(Refer Slide Time: 25:19)

$n = 20, k = 0, p = 0.01$

$$f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$
$$= 0.82$$

82% chance that a sample of 20 individuals will fail to find any infections

Let us cross check with free Graph pad online software
<http://www.graphpad.com/quickcalcs/>



Now we can also check with the online software also.

(Refer Slide Time: 25:29)


$n = 20, k = 0, p = 0.01$

$$f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$
$$= 0.82$$

82% chance that a sample of 20 individuals will fail to find any infections

Let us cross check with free Graph pad online software
<http://www.graphpad.com/quickcalcs/>

	Number of Successes	Exact Probability	Cumulative Probability
Number of trials (or subjects) per experiment: 20	0	81.791%	81.791%
Probability of "success" in each trial or subject: 0.010	1	16.523%	98.314%
	2	1.586%	99.900%



Using the same online software, for example same thing for getting 0, it gives you 81 % or 82 %. If I want to find at least 1 % with that, 98 % will happen actually. Same thing we can do it using this. So, what we do.

(Refer Slide Time: 25:52)

The screenshot shows a presentation slide with the following content:

$n = 20, k = 0, p = 0.01$

$$f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$
$$= 0.82$$

82% chance that a sample of 20 individuals will fail to find any infections

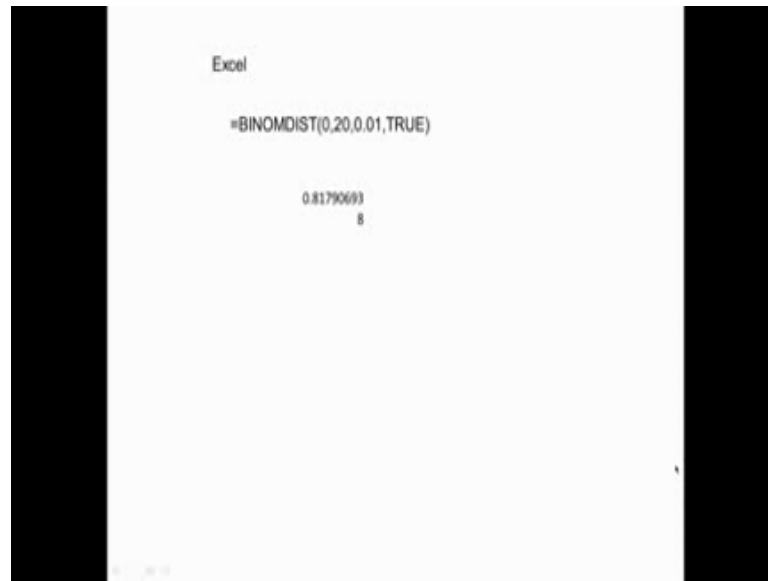
Let us cross check with free Graph pad online software

<http://www.graphpad.com/quickcalcs/>

Number of trials (or subjects) per experiment	Number of Successes	Event Probability	Cumulative Probability
20	0	0.81791%	0.81791%
	1	16.527%	0.83544%
	2	1.586%	0.85030%

We will go to the GraphPad and then I go back and I will put 20 then I will put 0.01 then calculate probability. As you can see here there is 81 % probability or 82 % probability that not a single number of successes 0, that means not a single person with that particular disease. Obviously my data is too little my sample size is too little that I may miss out. So, you must be very careful when you select sample, a very small sample can make you conclude wrongly. That sample size is a very very important parameter and we are going to talk about that in other cases also as we go along. So, with the very small sample size for example, here 20 people with the 1 % probability I may say that 82 % of the time there will not be even a single person infected with that disease in this sample of 20. So you can see that we can show it using this equation or we can go to that software GraphPad online and then get the same answer. Even with the Excel also we can do the same thing, we go to the Excel. We type BINOM distribution f x. We have BINOM distribution, we have number of successors we are talking about 0, trials is 20, probability is 0.01, then we can say false or true it does not matter, false then we get again you can see the answer is 82 %. So. 82 percent of the time we will not be finding any infected person, if I take a sample of only 20.

(Refer Slide Time: 28:03)



So you have to be very very careful on that, 82 % of the time we will miss out we will come to a wrong conclusion.

(Refer Slide Time: 28:11)

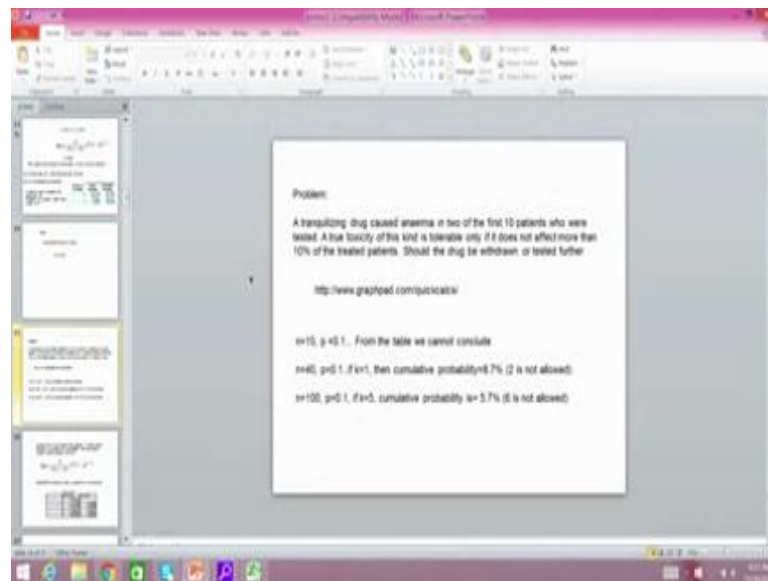


Let us look at another problem. A tranquilizing drug caused anemia in 2 of the first 10 patients who were tested. I took 10 patients and then I am I gave the drug first 2 patient

had some toxicity problem but then a true toxicity of this kind is tolerable only if it does not affect more than 10 % of the treated patient, but here the first 2 patients themselves had the problem. Should that drug be withdrawn or tested further. So only 10 % of the patients can have this type of toxicity affects but here with 10 patients, 2 of them are having problem. So, should the drug be taken out?

We are in big problem, so let us go for example.

(Refer Slide Time: 29:09)



The GraphPad, then we will say 10 patients and then we want to say 0.01 percent, calculate probability, that is very very high. So, we cannot conclude because it is showing almost very high probability almost 34 % whereas we want to have less than we want to have 10 % only. Whereas if I take a larger population, for example, if I take n equal to 40, if I take a larger population for testing and then I keep the same 10 %, when I calculate the probability then as you can see here, if I go to 2 % successes here it is still going to 14 % of probability. The cumulative if you look at it, it is coming to again 22 % whereas if you want to have less than 5 % as a possible number then obviously, if I go to say n = 100, if I go to n = 100. For example, suppose I take a sample of a 100 patients and then do the study as we can see here, out of the 100 patient I can have up to 5 patients having toxicity, I will be within that 10 percent limit but if I go beyond that I will have numbers going up.

Obviously what it means is the number of samples I have taken should be considerably large in order to prove that the toxicity is less than 10 %. Obviously in this particular case also we can see the sampling size has to be much larger.

(Refer Slide Time: 31:20)

Suppose 30% of the students wear glasses. If I take a random sample of 10 students find the probability that the number of students wearing glasses is at most 4.

$$f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

BINOMDIST(number_s, trials, probability_s, cumulative)

p=0.3, n=10		
k	exact prob	cum prob
0	0.0282475	
1	0.1210608	0.149308
2	0.2334744	0.382783
3	0.2668279	0.649611
4	0.2001209	0.849732

Suppose let us look at another problem 30 % of the students wear glasses. If I take a random sample of 10 students, find the probability that the number of students wearing glasses is at most 4? It is people of different types, you can have people wearing glasses, you may get no one, you may get 1 person wearing glasses, you may get 2 persons wearing glasses, you may get all the 4 person wearing glasses, right. We have a 30 % of students, here $p = 0.3$ and then you have $n = 10$ and then you want to look at various conditions of 1 person wearing, 2 person wearing, 3 person wearing, 4 person wearing glasses, that will be the k values. So, we can use this particular function.

(Refer Slide Time: 32:22)

Suppose 20% of the students wear glasses. If I take a random sample of 10 students find the probability that the number of students wearing glasses is atleast 4.

$$P(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

BINOMDIST(number_x, trials, probability_x, cumulative)

x	prob	cum prob
0	0.107374	0.107374
1	0.263893	0.371267
2	0.246032	0.637299
3	0.153520	0.790819
4	0.088080	0.878900

I take 10 students, the probability is 0.3. So I calculate the probabilities, as you can see here 0 person wearing glasses, 1 person wearing glasses, 2 person, 3 person and so on. 0 person wearing glasses will be 2.82 % but if you are talking about 1 person wearing glasses out of this 10 is 12 %. 4 persons wearing glasses is 20 % but if I add up all these, that means, if I take 10 students out of this lot, students wearing 1 or 2 or 3 or 4 person wearing glasses will be so many percent, 84 % or 0 glasses. So, this is the cumulative and this is the exact probability here. You can use this QuickCalcs of the GraphPad to identify the probability distribution function for a Binomial Distribution. You can use this equation or we can use the Excel function or we can use the GraphPad software also. So all these are possible to get, as you can see here this is the cumulative, this is the exact probability for 0 person wearing glasses, 1 person wearing glasses, 2 person, 3 person, 4 person like that you know it goes up to n of 10.

(Refer Slide Time: 34:21)

Disease with known mortality = 10%, what is the minimum number of patients required to demonstrate the efficacy of the completely curative drug

$\pi(\text{survival}) = 0.9$

$(1 - \pi)(\text{death}) = 0.1$

$0.9^n < 0.05$ (for 95 % confidence)

$N > 29$

<http://www.graphpad.com/quickcalcs/>

30, 0.9

NPTEL

Now, let us look at another problem. You know there is a disease with known mortality 10 %, what is the minimum number of patients required to demonstrate the efficacy of the completely curative drug? That means there is a disease of mortality of 10 % that means, 0.1, survival if you take as π 0.9 **1 - 5 is death is 0.1**. I want to show completely curative, that means, I do not want to see any disease. **If I take n patients and survival probability for each of the patient is 0.9, it will become $0.9 \times 0.9 \times 0.9^n$** . Now this should be less than 0.05 because why? 0.5 is 5 % that means that gives you 95 % confidence. Do you understand? Thus mortality is 10 percent that **is 0.1**, survival is 0.09. If I call 5 survival as 0.09, 1 - 5 death is equal to **0.1**.

Now, if I take n patients then survival for each one is **0.9. So, $0.9 \times 0.9 \times 0.9^n$, I do it n times that is why I have 0.9^n** . Now this should be less than to get a confidence of 95 %, this should be less than 0.05. So, if I calculate this from this n I get n should be greater than 29, that means, I should have at least 29 patients and show on all of them none of them die. If I do that then I have a 95 % confidence that drug has a completely curative affect.

This approach tells you how to select the number of subjects or number of samples in the in our problem. We looked at many different cases where we used Binomial Distribution and Binomial Distribution is based on successes when you take a sample of n. So k

successes in a sample of n and the probability of each one happening p it tells you, what is the probability of k successes in a sample of n , if the probability for each event is p and that is what is Binomial Distribution is all about. We can use it like, if there are 30 % of the students wear glasses in a class. If I take 10 students, what is the probability that 4 of them will be having glasses? If I have a disease which happens 2 % in India, if I take a family of 20 people in a house, how many of them will have this particular disease. So, for all these we use this Binomial Distribution very effectively and it is very very useful.

I also taught you how to use the binomial distribution using the **formula**.

$$f(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

We can do it numerically or we can use the Excel, all of us have Excel there is a function called Binom Distribution in the Excel where you can substitute it and calculate or you can use online software called GraphPad, I showed you the link to that software you can substitute the data and get the values. So, all these approaches are possible and you can see binomial distribution is very very useful in clinical trials and large data analysis.

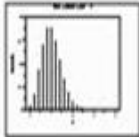
The next class we will look at something called the Poisson distribution again this is a Discrete Distribution which talks about events.

(Refer Slide Time: 38:16)


Discrete Distributions

[Poisson Distribution](#)

the probability of a number of independent events occurring in a fixed time.



If the probability p is small and the number of observations is large



Again, Poisson is an extension of Binomial Distribution.

Thank you very much.