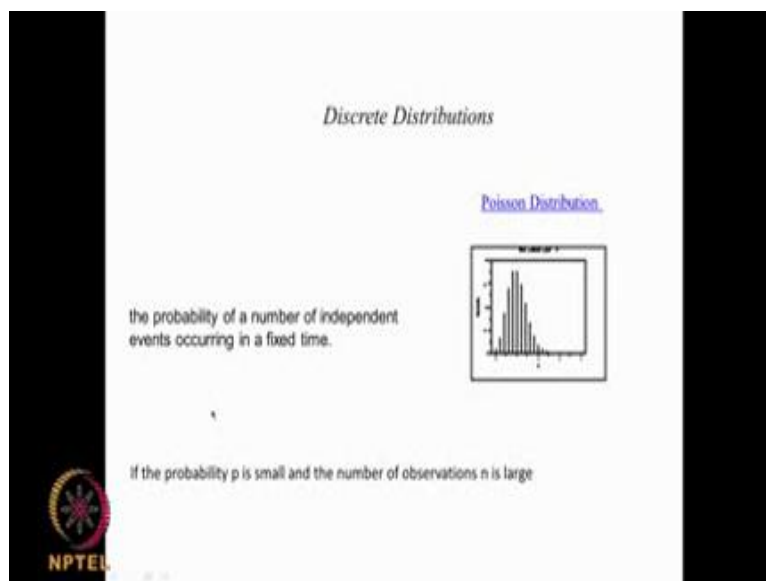


Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 04
Poisson Distribution

Welcome to the next class. Today we are going to talk about another distribution it is called Poisson distribution. The previous class we talked about the binomial distribution. In binomial distribution you have n samples and you have k successes and probability of each of the success is half, so you are expected to find out the total probability. Whereas in Poisson distribution, when n becomes very large then the binomial sort of tends into Poisson distribution and that is what it is all about actually.

(Refer Slide Time: 00:45)



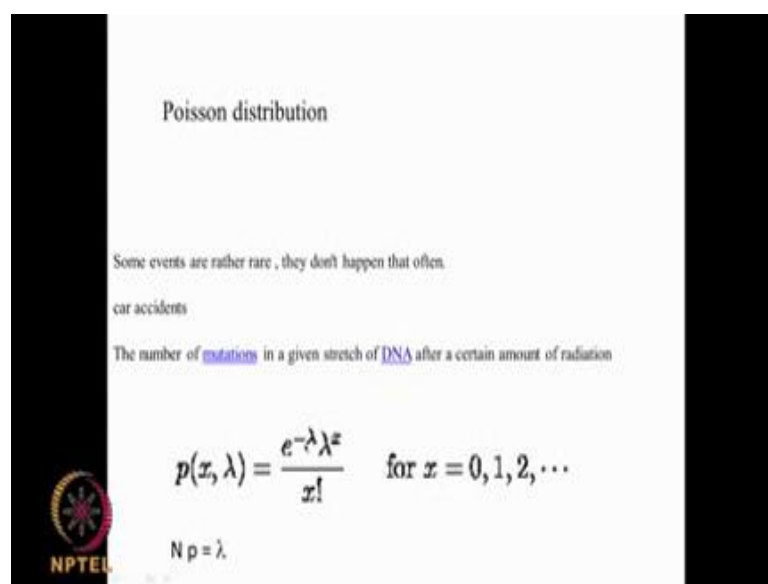
We have the probability of number of independent events occurring in a fixed time. The probability of a particular event occurring in a fixed time. For example, number of car accidents that is happening in a metro city in India in past one month or number of infant deaths in India in the past one year and so on. So, we are talking about based on a large number of data we are trying to find out the events actually. The probability p will be very small whereas the number of observation will be very large. It is sort of related to the

previous one which we saw the binomial but here the n is very, very large and p becomes very, very small actually. This is also very useful in biology as I am going to talk about a few examples. The Poisson distribution the equation looks like this,

$$p(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

x could be 0, 1, 2 and so on.

(Refer Slide Time: 01:41)



Poisson distribution

Some events are rather rare, they don't happen that often.

car accidents

The number of mutations in a given stretch of DNA after a certain amount of radiation

$$p(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

$Np = \lambda$

NPTEL

How do you know λ ? λ is given by this relation $n \times p = \lambda$,

here n is very, very large actually. Some events are rather rare, they do not happen often like car accidents or infant deaths and so on actually. For example, number of mutations in a given stretch of DNA after it is exposed to radiations, in such sort of situations we use Poisson distribution. The equation is like this,

$$\frac{e^{-\lambda} \lambda^x}{x!}$$

the λ is given by this relation n is the total data set, p this is a probability of occurrence. So, x could be 0 if you are looking at 0 events, x could be 1 if you are looking at 1 event, 2, 3 and so on actually.

(Refer Slide Time: 02:39)

Suppose the average number of fatalities due to car accidents in a city in India on any day is 5. What is the probability that fewer than four such fatalities will occur on any particular day?

Here $\lambda = 5$
 $x = 0, 1, 2, \text{ or } 3$; since we want to find the likelihood of fewer than 4 accidents

$$P[x < 4, 5] = [(e^{-5})(5^0) / 0!] + [(e^{-5})(5^1) / 1!] + [(e^{-5})(5^2) / 2!] + [(e^{-5})(5^3) / 3!]$$

$$= 0.2650$$

NPTEL

Let us look at an example, suppose the average number of fatalities due to car accidents in a city in India on any day is 5. This number one might have collected over a very long period of time. So, it is almost like a huge population data and then you get this data actually. It says on any particular day, there could be 5 accidents which lead to fatal result. What is the probability that fewer than 4 such fatalities will occur on any particular day? If I say tomorrow, what will be the probability that fewer than 4. That means it could be 0 accident, it could be 1 accident, it could be 2 accident, it could be 3 accident, because we are saying fewer than 4.

What is the probability? Say if tomorrow in that particular city, you have fewer than 4 accidents. What you do? You know this **equation,**

$$\frac{e^{-\lambda} \lambda^x}{x!}$$

and you need to put $x = 0$ then $x = 1$, $x = 2$, $x = 3$ then add up all of them. What will be the λ here? λ is 5; because we are talking about in any given day statistically they have found that there will be 5 accidents per day. So, λ will = 5, then x you put it as 0, then x you put it as 1, you put x is 2, then x is 3, then add up all that and that will give you the entire probability. So, probability that fewer than 4 accidents take place, that means $x < 4$, that means it can be $x = 0$ or $x = 1$, $x = 2$, $x = 3$ and λ is 5. So what you do,

$$e^{-\lambda} \lambda^x$$

In this case, it is 0 divided by 0! then $(e^{-5})(5^1) / 1!$, $(e^{-5})(5^2) / 2!$, $(e^{-5})(5^3) / 3!$ So, 0! is 1 and anything raise to the power 0 is also 1. When you do all these adding up you get 0.265. So, the probability of having fewer than 4 accidents say tomorrow or any particular day will be 26.5 percent; that is what it is.

(Refer Slide Time: 05:01)

Suppose the average number of fatalities due to car accidents in a city in India on any day is 5. What is the probability that fewer than four such fatalities will occur on any particular day?

Here $\lambda = 5$
 $x = 0, 1, 2, \text{ or } 3$; since we want to find the likelihood of fewer than 4 accidents

$$P[x < 4, 5] = [(e^{-5})(5^0) / 0!] + [(e^{-5})(5^1) / 1!] + [(e^{-5})(5^2) / 2!] + [(e^{-5})(5^3) / 3!]$$

$$= 0.2650$$

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

NPTEL

We can also check with the online software. Yesterday I introduced this software called Graph pad online software and this is the link for that software. We can do the same calculations with that software also.

(Refer Slide Time: 05:15)

Suppose the average number of fatalities due to car accidents in a city in India on any day is 5. What is the probability that fewer than four such fatalities will occur on any particular day?


Here $\lambda = 5$
 $x = 0, 1, 2, \text{ or } 3$; since we want to find the likelihood of fewer than 4 accidents

$$P(x < 4, 5) = \left[\frac{(e^{-5})(5^0)}{0!} \right] + \left[\frac{(e^{-5})(5^1)}{1!} \right] + \left[\frac{(e^{-5})(5^2)}{2!} \right] + \left[\frac{(e^{-5})(5^3)}{3!} \right]$$
$$= 0.2650$$

Let us cross check with free Graph pad online software

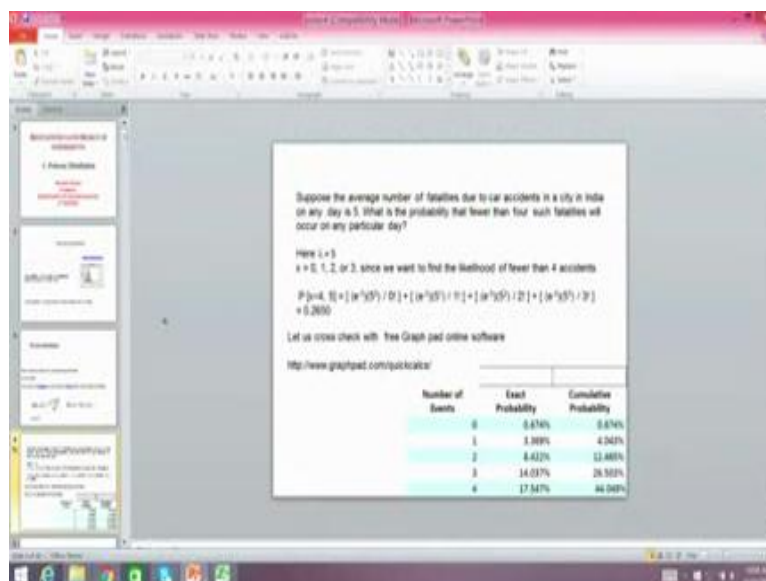
<http://www.graphpad.com/quickcalcs/>

Number of Events	Exact Probability	Cumulative Probability
0	0.674%	0.674%
1	3.369%	4.043%
2	8.422%	12.465%
3	14.037%	26.503%
4	17.547%	44.049%



I can substitute and you get 0, 1, 2, 3, 4, if you do the cumulative as you can see for 0 then 1 cumulative will be this plus this giving this, for 2 cumulative this plus, this plus, this, for 3 cumulative this plus, this plus, this plus, this that comes to 26.5 percent which is matching with 26.5. Shall we use the graph pad?

(Refer Slide Time: 05:42)



Suppose the average number of fatalities due to car accidents in a city in India on any day is 5. What is the probability that fewer than four such fatalities will occur on any particular day?

Here $\lambda = 5$
 $x = 0, 1, 2, \text{ or } 3$; since we want to find the likelihood of fewer than 4 accidents

$$P(x < 4, 5) = \left[\frac{(e^{-5})(5^0)}{0!} \right] + \left[\frac{(e^{-5})(5^1)}{1!} \right] + \left[\frac{(e^{-5})(5^2)}{2!} \right] + \left[\frac{(e^{-5})(5^3)}{3!} \right]$$
$$= 0.2650$$

Let us cross check with free Graph pad online software

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

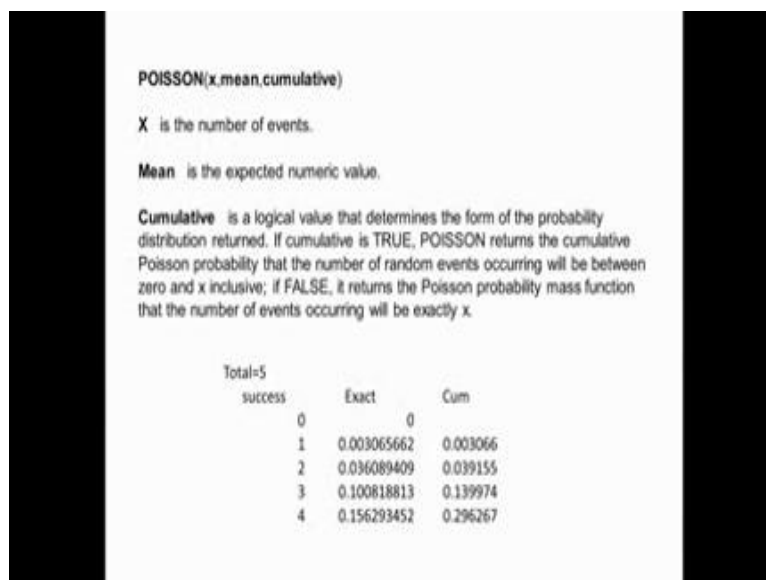
Number of Events	Exact Probability	Cumulative Probability
0	0.674%	0.674%
1	3.369%	4.043%
2	8.422%	12.465%
3	14.037%	26.503%
4	17.547%	44.049%

So, let us use the graph pad software. As you can see it tells you that we can use Binomial Poisson and so on. So, we use this then go forward then this is the one we again use this and

go forward. We have here it says Poisson distribution, so average number of objects that means on an average they have seen 5 fatalities on any particular day in that city. We put 5 and then we calculate the probability.

You see this, for 0 accidents on any particular day the probability will be 0.67 %, for only 1 accident the probability will be 3.36 %, but if it is 0 or 1 accident then you need to add these 2 that is cumulative, so it comes to 4 %. For 2 accidents, it will be 8.42 %, for 3 accidents it will be 14 %, but if you are talking about 0 or 1 or 2 or 3 accidents on any particular day, you need to add all these. So, that is the cumulative we get 26.5; so, we got the same answer. We can use this graph pad software also to do the same calculations. So, that is what we got here 26.5 % for less than fewer than 4 or less than 4 accidents.

(Refer Slide Time: 07:17)

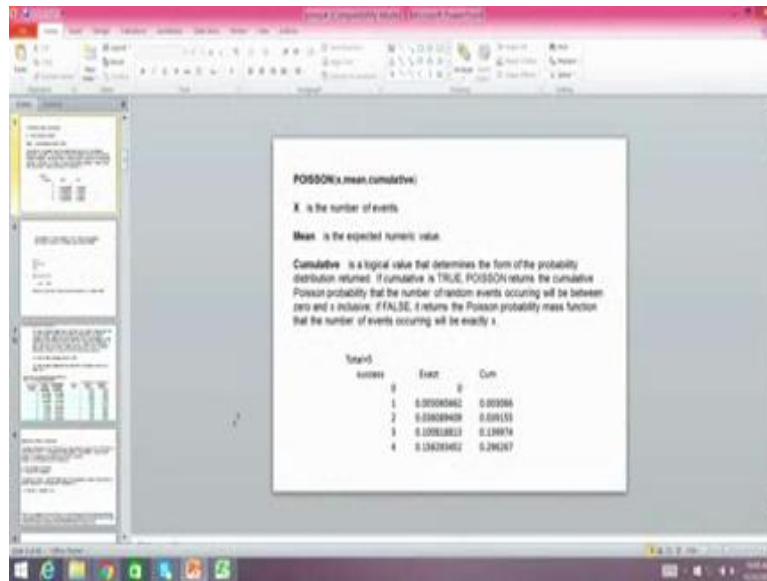


The screenshot shows a Poisson distribution calculator interface. It includes a title 'POISSON(x,mean,cumulative)', a description of 'X' as the number of events, and 'Mean' as the expected numeric value. A detailed description of the 'Cumulative' parameter is provided. Below this is a table with the following data:

Total=5 success	Exact	Cum
0	0	0.003066
1	0.003065662	0.003066
2	0.036089409	0.039155
3	0.100818813	0.139974
4	0.156293452	0.296267

Even Excel has this option we have something called function called Poisson. x mean cumulative, x is the number of events, x is the number of events, m is the mean that means, m is the expected numerical value cumulative true or false. Like in binomial, if you put false you will get the exact value, whereas if it is true you will get the cumulative value. We put total as 5 and we want to look at 0, 1, 2, 3, 4, we can use the same function here in excel also. Let me do that, we go to the Excel.

(Refer Slide Time: 08:01)



We have the statistical then we have the Poisson, so you see the Poisson distribution. When we say Poisson distribution x is the number of events, mean is the expected numerical value and here we put false to get the absolute or we can put true to get the cumulative. So, x is the number of events 5 for example, I want to look at say 3 and then I say true and then I get this, it gives you something here which is not correct, we got 26 percent. In the Poisson distribution, we have x is the number of events we are looking at, mean is the expected numerical value and cumulative is true in this case. Here we are having 5 fatalities expected on any particular day but we are looking at minimum of less than 4, so I have put 3 here and I have put true here and we get the answer as 26.5, which matches with whatever we got here. We got 26.5 and 26.5 using different method.

So, we can use this formula or we can use this graph pad software or we can use this Poisson distribution function that is available in Excel as well actually. Here we say x is the number of events we are looking at, it could be 1, 2, 3, 0 and then is the total and then cumulative could be true or false. Let us look at another problem where Poisson distribution is useful.

(Refer Slide Time: 10:42)

The probability of a birth defect is 10%. What is the probability that no one in a family of 10 people have that birth defect

$Np = \lambda$
 $10 * 0.1 = \lambda$
 $x=0$

$P[0,1] = e^{-\lambda} / 0!$
 $= e^{-1} = 0.367$

Probability of at least 1 person with the birth defect = $1 - 0.367 = 0.633$

There are probability of a birth defect is 10 % this is data which may be collected over a very very long period of time. So, the probability of a birth defect is 10 %. What is the probability that no one in a family of 10 people have the birth defects? I have a family in a village there are 10 people, what is the probability that no one in that family will have this birth defect? But the probability of birth defect is 10 %. So how do you do this? Again we need to know λ here, $n p = \lambda$ here n is 10 people and the p probability is 0.1. So λ comes out to be 1. Here x we want to be 0 because we do not want to have any birth defect here. The $P[0,1]$ will be

$$P[0,1] = e^{-1} / 0!$$

that comes to e^{-1} which is 0.367 that means there is 36 % probability that in a family of 10 people nobody is having the birth defect. We can also calculate 1 person having the birth defect we use at least 1 person having the birth defect then we can look at putting different numbers here based on what is the x we are looking at actually. Now if you want to say probability of at least 1 person with the birth defect that means the birth defect could be all 10 having birth defect, all nine having birth defect, all 8, all 7, all 6 having birth defect, 5 having birth defect 4, 3, 2, 1. So, it will be like 1 minus nobody having birth defect, so that is why we have 1 minus 0.367 that is 0.633 that means probability of at least one person having birth defect in that family of 10 is 63 %. Here at least 1 person means 1 could be having birth

defect, 2 could be having, 3 or 4 or 5. So it is exactly 1 minus of nobody having the birth defect.


(Refer Slide Time: 13:02)

you were to scatter seeds over a vast field from a plane. Imagine also that you have divided the field up into blocks of equal size. (you haven't dropped a trillion seeds, just a few thousand), and if this probability is the same everywhere across the entire field, and if seeds are independent of each other then the number of seeds per block should follow a Poisson distribution: (ref: <http://www.zoology.ubc.ca/~bio300b/poissonnotes.html>)

you want at least one seed per plot, then?

you want at least 2 seeds per plot (then 40% of the area will have 0 or 1 seed only) !

Number of Events	Exact Probability	Cumulative Probability	Number of Events	Exact Probability	Cumulative Probability
0	36.788%	36.788%	0	33.534%	33.534%
1	36.788%	73.576%	1	27.067%	40.601%
2	18.394%	91.970%	2	18.045%	58.712%
3	6.131%	98.101%	3	9.022%	64.735%
4	1.533%	99.634%	4	3.609%	68.344%
5	0.307%	99.941%	5	1.203%	69.547%
6	0.051%	99.992%	6	0.344%	69.890%
			7	0.086%	69.976%
			8	0.018%	69.995%



Now let us look at another interesting problem. This I took it from website here you were to scatter seeds over a large field from plane. Imagine that you have divided the field into blocks of equal size; you have not dropped millions and trillions of seeds but only small amount of seed. What is the probability that the seeds are independent of each other? Of course 1 seed settling down is not going to effect the other seeds action. So what is the probability that may be at least 1 seed you get per plot? Or what is the probability at least 2 seeds you get per plot? We can again use your Poisson distribution. As you can see, we can say at least 1 seed we can have about 36 percent, at least 2 seeds per plot you can have 40 percent that means 2 seeds 0 or 1 seed it could be. So like that we can calculate using the graph-pad software.

(Refer Slide Time: 14:22)

average number of blocks per unit volume is seeds per unit time.

you were to scatter seeds over a vast field from a plane. Imagine also that you have divided the field up into blocks of equal size. (you haven't dropped a trillion seeds, just a few thousand), and if this probability is the same everywhere across the entire field, and if seeds are independent of each other then the number of seeds per block should follow a Poisson distribution: (ref: <http://www.zoology.ubc.ca/~bio300b/poissonnotes.html>)

you want at least one seed per plot, then?

you want at least 2 seeds per plot (then 40% of the area will have 0 or 1 seed only) !

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

Number of Events	Exact Probability	Cumulative Probability	Number of Events	Exact Probability	Cumulative Probability
0	36.788%	36.788%	0	13.534%	13.534%
1	36.788%	73.576%	1	27.067%	40.601%
2	18.394%	91.970%	2	27.067%	67.668%
3	6.131%	98.101%	3	18.045%	85.713%
4	1.533%	99.634%	4	9.022%	94.735%
5	0.307%	99.941%	5	3.609%	98.344%
6	0.051%	99.992%	6	1.203%	99.547%
			7	0.344%	99.891%
			8	0.086%	99.976%
			9	0.019%	99.995%



Now let us look at another problem.

(Refer Slide Time: 14:23)

Restriction Sites in a Genome

The base composition of the *Thermococcus celer* genome is about 0.21:0.29:0.29:0.21 (mole ratio A:C:G:T). If the sequence were random, the probability that any given position in the genome is a *SpeI* site (ACTAGT) would be

$$P(\text{SpeI}) = (0.21)(0.29)(0.21)(0.21)(0.29)(0.21)$$

$$= (0.21)^4(0.29)^2 = 0.000164$$

= 1 site per 6100 basepairs.

The genome is about 1,890,000 base pairs, so the expected number of *SpeI* sites in a random sequence of this length and composition is

$$\lambda = 0.000164 \times 1890000 = 310.$$

https://www.google.co.in/url?url=https://www.life.illinois.edu/mcb/432/Handouts/Binomial_and_Poisson.pdf&rct=j&frm=1&q=&esrc=s&sa=U&ved=0ahUKEwjKh97NfPbJAHWTB4KHcpkDQIQfgxMAU&usq=AFQjCNFe4bsr8oIPbdfpNH3Jcg_xuxquQ



This is something related to Genome. The base composition of the *Thermococcus celer* genome is about 0.21 is to 0.29 is to 0.29 is to 0.21 that is the mole ratio A C G T. The probability will be of a or c or g or t will be in this ratio actually, if the sequence were

random the probability that any given position in the genome is a Spel site that is ACTAGT would be 0.21 of because A is 0.21, C is 0.29, T is again 0.21 and A is 0.21, G is 0.29 and T is 0.21. So the probability that you have a genome sequence Spel site ACTAGT will be all these actually that is equal to 0.000164, that is one site per 6100 base pairs. Now this genome is about 1890000 base pair, so the expected number of Spel sites in a random sequence of this length and composition will be this 0.00164 multiplied by this, we get a $\lambda =$ to 310 we can substitute that into our equation for less than 5.

(Refer Slide Time: 16:02)

But The observed number of sites is 5.13 The probability of 5 or fewer sites when the expected number is 310 is:

$$P(x \leq 5) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5)$$

$$= (1 + 310 + 48050 + 4965167 + 384800417 + 23857625833) e^{-310}$$

$$= 24247439778 e^{-310}$$

$$= 5.7 \times 10^{-125}$$

This is a small probability for a random event. Therefore it is reasonable to reject the model that the nucleotide sequence of the *Thermococcus celer* genome is random with respect to this sequence.

NPTEL

We want to look at less than 5, $P = 0, 1, 2, 3, 4, 5$ so we are looking at x is equal to 0 λ is equal to 310, $x = 1 \lambda = 310$, $x = 2, \lambda = 310$, $x = 3 \lambda = 310$, $x = 4 \lambda = 310$, $x = 5 \lambda = 310$.

So if we substitute all these now we end up with such a very very small number, 5.7×10^{-125} but what is observed the observed number of sites is 5.13 which is very big. Obviously, it is not a random event because if it is a random event you should get this as a probability but actually you are observing 5.13 that is, at least 5 or fewer sites therefore it is reasonable to reject the model that the nucleotide sequence of the *Thermococcus celer* genome is random with respect to the sequence. So it is not a randomly happening because if it has to happen randomly the probability of that is this but actually you observe almost 5.13 or 5 or fewer sites, which is a large number so this sequence of ACTAGT which is a Spel site happening is not a random event. It is very interesting problem this it was taken from this particular site

and we can do similar studies on genome sequences and when you see a particular sequence you can see whether it is in random event using Poisson distribution or it is not a random event.

(Refer Slide Time: 17:59)

Poisson Distribution

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

Poisson probability is: $P(x; \lambda) = \frac{e^{-\lambda} (\lambda^x)}{x!}$


Probability of 0 event = $e^{-\lambda} [\lambda^0 / 0!]$
 Probability of 1 event = $e^{-\lambda} [\lambda^1 / 1!]$
 Probability of 2 event = $e^{-\lambda} [\lambda^2 / 2!]$
 Probability of 3 event = $e^{-\lambda} [\lambda^3 / 3!]$

where x is the actual number of successes that result from the experiment
 average number of successes is λ .

Confidence interval for the average count, L .

$\lambda = L \pm t \sqrt{L}$

$t = 1.96$ for 95% confidence interval and 2.58 for 99% confidence interval



$$\lambda = L \pm t \sqrt{L}$$

And once again to recap Poisson distribution you have when n is very large, so you have a something called λ here which is the governing term λ is given by $n \times p$, p is the probability, n is the number of total number of events, it is given by

$$\frac{e^{-\lambda} \lambda^x}{x!}$$

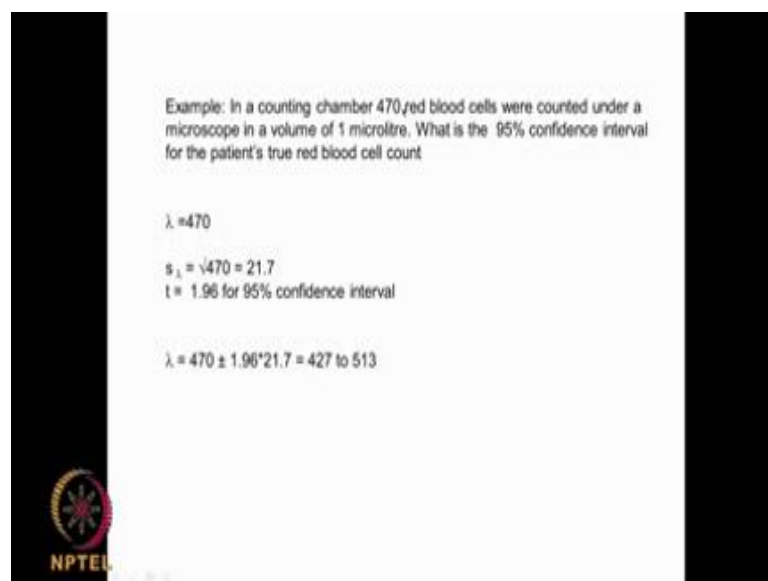
If I want to look at 0 event, then I put $x = 0$ here, if you want to look at 1 event I put one here, if I want two I put 2 here 3, 3 and if I am looking at either absolute or I can even look at cumulative that means probability of fewer than 4 events, if I say then I need to add all these. Now we can use Poisson distribution also to find out confidence interval on the count for example, if I am counting number of bacteria colonies in a plate or if I am counting red blood corpuscles in the blood these are all individual events.

Obviously, if I take a different blood sample, I may get a different number, if I take a different blood sample from the same patient I may get a different number. Obviously, there will be a range it cannot be an absolute single value that confidence interval is given by this term plus

or minus $t \sqrt{I}$. I is the average count but there is a confidence interval associated with this I which is given by plus or minus $t \sqrt{I}$, where t is 1.96 for 95 percent confidence interval and 2.58 for 99 % confidence interval. This equation is very useful. Like I said if I am counting the number of live bacterial cells in my plate I get a number say 10^{20} .

What is the range? What is the confidence interval? If I want to know if I am looking at the red blood corpuscles of a volunteer, when I take a sample and count I may get some number when I take another sample I may get another number, like that if I keep on doing it I may get large set of numbers. Obviously there will be a confidence interval that is given by this particular term $t \sqrt{I}$, I is the count and t is given by 1.96 for a 95 % confidence and 2.58 for a 99 % confidence interval. Let us look at some examples now that will give you.

(Refer Slide Time: 20:36)



Example: In a counting chamber 470 red blood cells were counted under a microscope in a volume of 1 microlitre. What is the 95% confidence interval for the patient's true red blood cell count

$$\lambda = 470$$
$$s_{\lambda} = \sqrt{470} = 21.7$$
$$t = 1.96 \text{ for } 95\% \text{ confidence interval}$$
$$\lambda = 470 \pm 1.96 * 21.7 = 427 \text{ to } 513$$

NPTEL

In a counting chamber, I have got 470 red blood cells counted under a microscope in a volume of one micro liter. So what is the 95 % confidence interval for the patients true red blood cell count? λ is 470; $\sqrt{470} = 21.7$, t is 1.96 for 95 % confidence, so 470 plus or minus this. Although we measure as 470 red blood cells, in reality if you want to mention it as a 95

% confidence it will vary between 427 and 538. For a 95 % I put this number as 1.96, whereas if it is a 99 % I put the number as 2.58. The true value will be 95 % of the time between 427 and 513.

(Refer Slide Time: 21:33)

Example: 100 agar plates containing antibiotics were streaked with 1 million bacteria each to determine the incidence of antibiotic mutants after incubation in all 58 mutant colonies were found.

Calculate the probability of finding 0 or 1 mutant colony per plate

For 0 mutant

$$\lambda = 58/100 = 0.58$$

$$e^{-\lambda} [\lambda^0 / 0!] = e^{-0.58} = 0.56$$

For 1 mutant

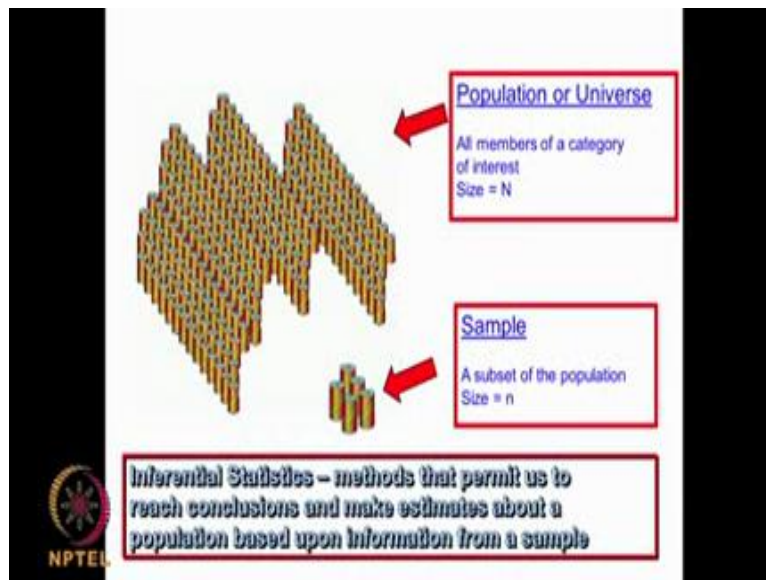
$$e^{-\lambda} [\lambda^1 / 1!] = e^{-0.58} * 0.58 = 0.325$$

Now let us look at another plate. There are 100 agar plates containing antibiotics were streaked with 1 million bacteria each to determine the incidence of antibiotic mutants after incubation. In all 58 mutant colonies were found, there were 58 mutant colonies. Calculate the probability of finding 0 or 1 mutant colony per plate? Obviously, if I want to find 0 now I found 58 colonies in 100 plates, my λ will be equal to 0.58 that is the incident. If I want to see 0 then I put x as 0. $[e^{-\lambda} / 0!]$ that is 56 %.

If I want to see one mutant colony, obviously, I will put $e^{-\lambda} [\lambda^1 / 1!]$, that gives you 32.5 %.

You can see is very very useful Poisson distribution we can use it for calculating events based on a probability. These events are independent of each other they are not related to each other. We can use Poisson distribution for getting a confidence interval for a count like I showed you in example on red blood corpuscle or if it is a bacterial colony I am counting. So Poisson distribution is very useful and λ is only factor which I need to know here lambda is it given by n p. Now I want to slightly switch gears and talk about something called population and sample.

(Refer Slide Time: 23:13)



Population is something which is very very big for example, when I say there are 5 fatalities on the road in metropolitan city that means this data must have been collected over a very very long time. So it is not that every day it will be happen 5 but it is collected over a very long time and that is called a population. Now if I am saying that the average height of Indians is 5 feet 5 inches so this data is collected over a very very large data set called population. If you say the number of defects children born will be 1 in 1 million in India then this data collected over a long period of time with large data set and that we can call it population and that is generally denoted like a capital **N**, whereas I may take a small sample that is called a subset of this population. I can because I cannot actually get all the population, but I can get a small sample. Suppose I am running a bolt factory, I take a few bolts and check their diameter and see whether it conforms with what I claim or if I take 10 people in Chennai and find their height and I will try to see whether it matches with the Indian height average of 5 feet 5 inches but that is a sample that is a very small number n. Even if I take 100 people even if I take 1000 volunteers in Chennai and measure their heights, still I would call it a sample, it cannot be a population.

There is always a population which is large, which is like a universe you know like the height of people in India that is very big whereas when I take a small sample that is denoted by small n and in statistics based on the results of sample we tried to predict what could be the

population whether the sample really falls into the population. I take 5 bolts and measure their diameter; I may get say 20.1 mm, 19.9 mm, 20 mm so I will take an average. Now I want to know whether this average confirms with that 20 mm bolts size which I have mentioned in my catalog. Is it really very close to 20 mm? Or is it very far away from 20 mm? When I take a sample, sample is always very small and when I take an average of that sample it will not be exactly 20 mm the average may be 19.5 or 20.5. Now this 20.5 is it very different from what I claim 20. Can I say they are same with 95 % confidence? Or can you say that they are same with 99 % confidence? That is what statistics is all about and so the concept of population and the concept of sample play very important role.

So sample is always very small whereas population is very very large. We can always collect samples and based on the sample results we tried to say whether it comes from the same population or whether it is not coming from the same population.

(Refer Slide Time: 26:57)

Descriptive Statistics - Central Tendency

Mean (Average Value)

Population Mean

Population size = N
Measure X for each member of the population

μ = Population Mean = Add up all the X-values in the population and divide by N

Sample Mean = $\frac{492.6 + 502.9 + 504.1 + \dots + 503.8}{20} = 501.9$

Sample Size = n = 20 for our example
Measure X for each member of the sample

X = Sample Mean = Add up all the X-values in the sample and divide by n

The Sample Mean is an estimate of the true Population Mean

NPTEL

When we have things like in continuous data, when I am measuring temperature 30.1, 30.2, 30.3 and so on, I can calculate something called mean. Mean is nothing but average. Everybody knows how to calculate mean, you add up all of them divided by the number of samples and then you get the mean.

Normally for population mean we represent it as μ bar whereas for the sample mean we may represent it as \bar{X} bar. Population mean we always represent it as μ here, whereas for sample mean, we generally represent it as \bar{X} bar. Like N is the population size whereas small n is the sample size. So always sample means are represented by \bar{X} bar whereas population is always represented by population mean is represented by μ here. Now this sample mean is an estimate of the true population mean, like I said, I take 10 bolts and then measure their diameter take an average that is \bar{X} bar, now μ is what is the real population mean which I say the bolts in my factory are 20 mm of size. Now this \bar{X} bar how close is it with μ , can I say that \bar{X} bar is a good representation of μ or is \bar{X} bar very far away from μ and so on and that is what statistical analysis is all about actually. Does \bar{X} bar very close to μ that I can say yes \bar{X} bar is a representation of the population or \bar{X} bar is not close to μ . It is not representation of this population. Now when you say close or not so close we use certain statistical terminology is like confidence limits, 95 % confidence, 99 % confidence and so on actually. So we will talk about all these much more in detail as we go along. Now median what is median?

(Refer Slide Time: 29:08)

Descriptive Statistics - Central Tendency

Median (Population and Sample)

The middle value when the data is placed in an ordered sequence

- For an odd number of data points, the Median = the middle value
- For an even number of data points the Median lies between the two data points in the middle

481.4 482.4 483.1 485.8 487.1 487.5 488.7 501.4 501.5 502.4 502.4 502.9 503.8 504.5 504.5 505.1 505.1 505.5 506.8 508.1 508.8 508.8

502.75

NPTEL

Median is the middle point of the data set. So if I have odd data sets median will be the exactly the middle point whereas if I have the even data set even means I have 20 numbers then, obviously the median will lie between the two data points in the middle actually.

Whereas if I have seen 19 so I may have here both sides 9 and 9 and the middle point will be in the center whereas if I have 20 I have 10 and 10 so obviously, the median will be the average of the 10th and the 11th point that is called median. So median is a middle point whereas mean is an average.

(Refer Slide Time: 29:54)

The mode is the value that appears most often in a set of data

3, 5, 7, 12, 13, 14, 20, 23, 23, 23, 23, 29, 39, 40, 56

23 is the mode

1, 3, 3, 3, 4, 4, 6, 6, 6, 9

3 and 6 modes (bimodal)

NPTEL

Then there is something called mode. Mode is the value that appears most often in data sets. If we have say data set here like this 23 is appearing many times so 23 is the mode of this data set. Now if the data set is like this you have 3 and 6 appearing, right? So you have 2 modes this is called a bi model distribution whereas this is a mono model. So we have a 3 terms the mean, median, mode. Mean is the average, median is the central point, mode is the value that appears most often in a set of a data. We will be using these in our statistical calculations and as we go along actually.

Thank you very much for your time.