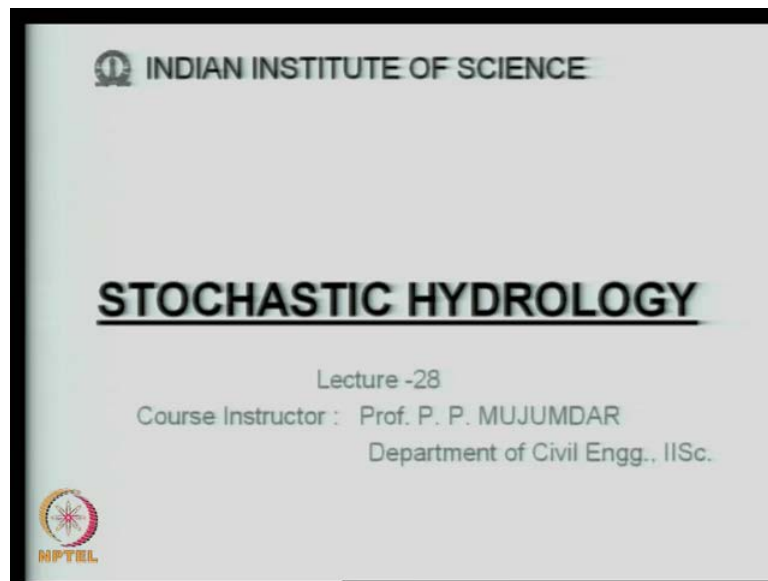


Stochastic Hydrology
Prof. P.P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

Lecture No. # 28
Goodness of Fit

(Refer Slide Time: 00:21)

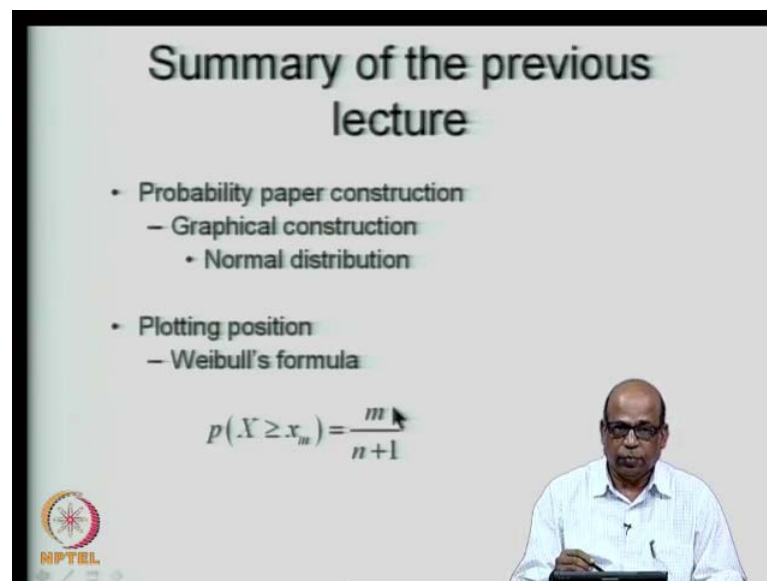


Good morning and welcome to this, the lecture number twenty-eight of the course stochastic hydrology. If you recall, in the previous lecture, that is, the lecture number 27, we discussed, essentially, the construction of probability paper, specifically the normal distribution paper, which we construct by graphical method. And in the lecture previous to that, we also discussed construction of probability paper by analytical methods in which we consider the exponential distribution, and then constructed the graphical probability paper.

What is it that we use the probability papers for? Let us say, that you have the normal distribution probability paper and you want to examine, whether the observed data, that you have, let us say stream flow at a particular location or rainfall seasonal rainfall at a particular location and so on, the hydrologic data you would like to examine, whether that fits normal distribution or not.

If you plot the data using the plotting positions on the normal distribution's probability paper and if it plots as a straight line, then at least you know it should plot as nearly a straight line, it should be acceptable as a straight line, then you can reasonably assume, that yes, it fits the normal distribution. What we will do in today's lecture is that we will carry this further and then look at some statistical tests, that are available also for examining, whether the sample data that you have observed data can be assumed to follow a particular specific distribution or not.

(Refer Slide Time: 02:58)



The slide is titled "Summary of the previous lecture". It contains the following content:

- Probability paper construction
 - Graphical construction
 - Normal distribution
- Plotting position
 - Weibull's formula

$$p(X \geq x_m) = \frac{m}{n+1}$$

In the bottom left corner, there is a logo for NPTEL (National Programme on Technology Enhanced Learning). In the bottom right corner, a man in a white shirt is visible, likely the presenter.

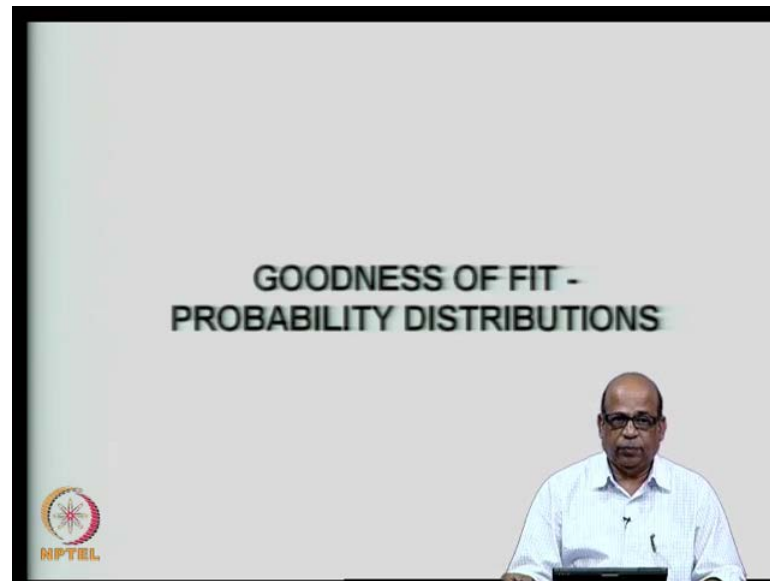
So, in the previous lecture we discussed the graphical construction of the probability paper and it is a, specifically the normal distribution is what we discussed, and we also introduced the plotting position, that we used for plotting the data on the probability paper and I enumerated a number of plotting position formulae. But we also saw in the previous lecture, that the Weibull's formula, which is given by p is equal to m divided by n plus 1, that gives you, that is the most widely used formula in hydrology, where m is the rank. When your, when you observe, when you arrange the data in descending order and assign the highest rank number 1 to the largest value and n is the number of values, so m divided by n plus 1, this gives you the plotting position for the particular value X_n .

We use the plotting position and then plot the data on the probability paper and then examine, whether it can be approximated as a straight line. If we, if we can, then we can reasonably assume, that the data comes from a sample, that follows normal distribution,

comes from the population, the sample data comes from a population, that follows normal distribution.

We now introduce specific statistical test, that we can carry, carry on, carry out for examining whether the sample data, that we have, follows a specific distribution or not.

(Refer Slide Time: 03:58)



Now, these are called as a test for goodness of fit, that is, we are examining whether the data can be fit to a particular distribution and how well we can fit, that data to a given distribution. So, it is called as a test for goodness of fit, for fitting probability distributions.

What I mean by that? If you have observed stream-flow data at a particular location and then you want to do all the analysis that we have discussed earlier. So, the first step is that you would like to fit here a particular probability distribution to that data. So, how well this data fits a specific distribution is a question that we would like to ask.

Let us say, that you are considering a candidate set of probability distribution. So, let us say normal distribution, log normal distribution, gamma distribution, etcetera. So, one by one you would like to try, does it fit normal distribution, does it fit log normal distribution and so on, so that you can use these distributions for your further analysis.

(Refer Slide Time: 05:16)

The slide is titled "Tests for Goodness of Fit". It contains a bulleted list with two main points and two sub-points. The first point is "Two ways of testing whether or not a particular distribution adequately fits a set of observations." The sub-points are "using probability paper." and "compare the observed relative frequency with theoretical relative frequency." To the right of the text is a graph showing a probability distribution curve. The y-axis is labeled "Probability" and ranges from 0.00 to 1.00. The x-axis is labeled "Area" and ranges from 0.00 to 1.00. A blue curve is plotted, which is nearly straight in the middle but curves away from a straight line at the extremes. In the bottom left corner of the slide, there is an MPTEL logo. A man in a white shirt and glasses is visible in the bottom right corner of the slide, appearing to be presenting.

- Two ways of testing whether or not a particular distribution adequately fits a set of observations.
 - using probability paper.
 - compare the observed relative frequency with theoretical relative frequency.

There are two ways of doing this, one of the ways is, as I mentioned in the previous lectures, use the probability distribution, probability papers. You, or construct, let us say, that you want to examine for normal distribution. You construct the normal probability distribution, which are also commercially available. You can plot the data on the commercially available normal distribution probability paper and then, plot the data as is shown here.

This is a data, this is the same example that I considered in the last lecture, you plot the data using the probability position. So, these are the probability positions, as you obtain from the Weibull's formula and then, if you can approximate this as a straight line, that is, this plots as a straight line, nearly as a straight line, then you can say, that it follows normal probability distribution.

In fact, as you can see from this example, if you are primarily interested in the central zone, central region like this, it fits the normal distribution fairly well only in the tail regions, either on the high extreme or on the low extreme. It departs significantly from the straight line indicating, that if your interest, in fact, lies on the tail regions, that is, either on the high extreme or on the low extreme, then you should not use from this result. At least, it indicates, that it should not use the normal distribution. So, that is the way we use the probability paper to examine whether the sample data indicates, that it

comes from the normal distribution or any other distribution for which the probability paper can be constructed or is available.

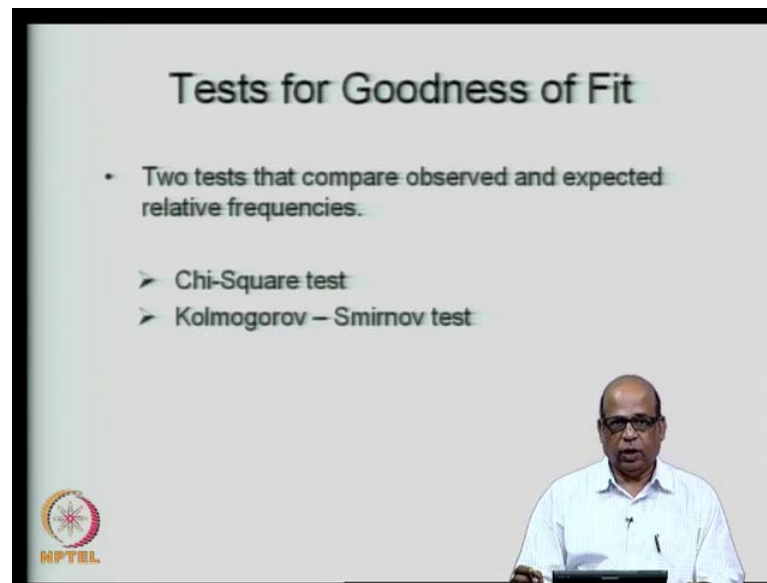
The second method is we can compare the observed relative frequency with the theoretical relative frequency arising out of that particular distribution, what do I mean by that? You have the observed data, so let us say, that you have a means of calculating the observed relative frequencies. What do I mean? The relative frequency, simply n_i by n , let us say we put it as p_i , which is the probability or the relative frequency n_i by n where n_i is a number of values, that falls in some class interval i and n is the total number of value. So, this is the relative frequency, the interpretation of probability in our earlier lecture where we have seen. In fact, it is a relative frequency. So, from the data you have an observed relative function n_i by n .

If the observed data were to fit a particular distribution, then this should be in some statistical sense close to the relative frequency as obtained by, as obtained theoretically or analytically for that particular distribution. So, that is what we mean by a theoretical relative frequency. So, we must have a means of calculating the relative frequency for this particular class interval, from the theoretical distribution, let us say, normal distribution.

So, we, for each of the class intervals we should calculate the relative frequency arising out of use of the normal distribution. So, you have on one hand the observed relative frequency on the other you have the theoretical relative frequency or the relative frequencies arising out of the particular distribution, which are also call it as expected relative frequency. By comparing these two, the observed relative frequency with the theoretical relative frequencies, we can conclude in some statistical sense with some confidence, some degree of confidence, that the data that we have can be approximated to follow a particular distribution.

Specifically, we introduce in today's lecture two tests, that are available, two classical tests, that are available for doing this. So, test, doing this analysis, these are the chi square test and the Kolmogorov-Smirnov test, both of which use the observed relative frequency and the expected relative frequency

(Refer Slide Time: 10:42)



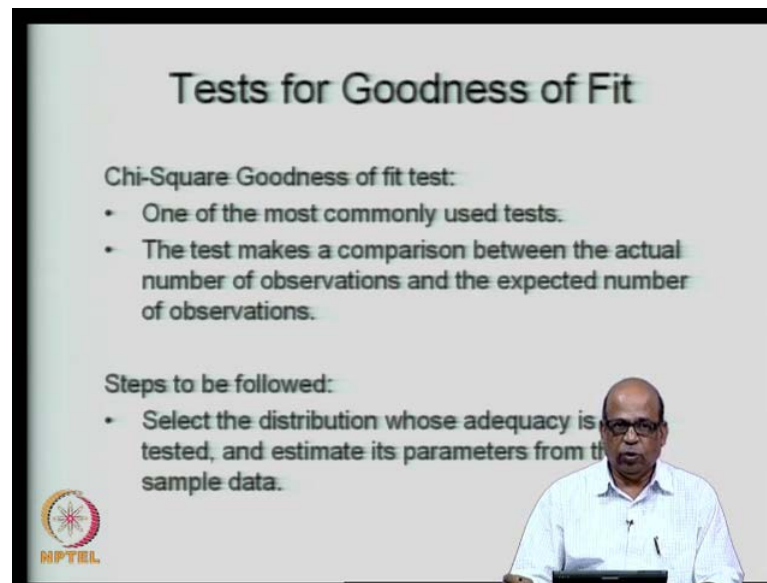
So, we will start with the chi square test and then go through some details of the chi square test and also we will examine through some example, some numerical example, how we actually apply the chi square test. Remember the question, that we are asking here is that we have a sample given sample data, specifically on stream flows rainfall, evapo-transpiration, soil moisture, etcetera, which we would like to use in further probabilistic analysis in our models or in our any applications for which we want to be sure or reasonably confident, that we can assume this sample data to have come from a given specific distribution, for example, normal distribution or log normal distribution, etcetera. So, that is the purpose.

Now, in the chi square what we do is. So, as I said, we will discuss these two tests. Now, chi square test and Kolmogorov - Smirnov test.

In the chi square test we actually compute or count the observed frequencies and compute the expected frequency and take the difference and look at what kind of departures we are getting of, on the observed frequency with respect to the expected relative frequency in some statistical sense.

In the Kolmogorov-Smirnov test, what we do is we compute these relative and observed frequencies and take the maximum departure and based on the maximum departure we conclude, whether this can be fit to a particular distribution or not. So, these are the two major tests that we frequently use. We will also discuss some limitations as we progress.

(Refer Slide Time: 11:42)



The slide is titled "Tests for Goodness of Fit". It contains the following text:

Chi-Square Goodness of fit test:

- One of the most commonly used tests.
- The test makes a comparison between the actual number of observations and the expected number of observations.

Steps to be followed:

- Select the distribution whose adequacy is tested, and estimate its parameters from the sample data.

In the bottom right corner of the slide, there is a small inset image of a man in a light blue shirt speaking at a podium. In the bottom left corner, there is a logo for NPTEL (National Programme on Technology Enhanced Learning) featuring a stylized sun or starburst design.

So, we will start with the chi square test. This, in most hydrologic application chi square test is very frequently used. As I said, it makes a comparison between the actual number of observations and the expected number of observations.

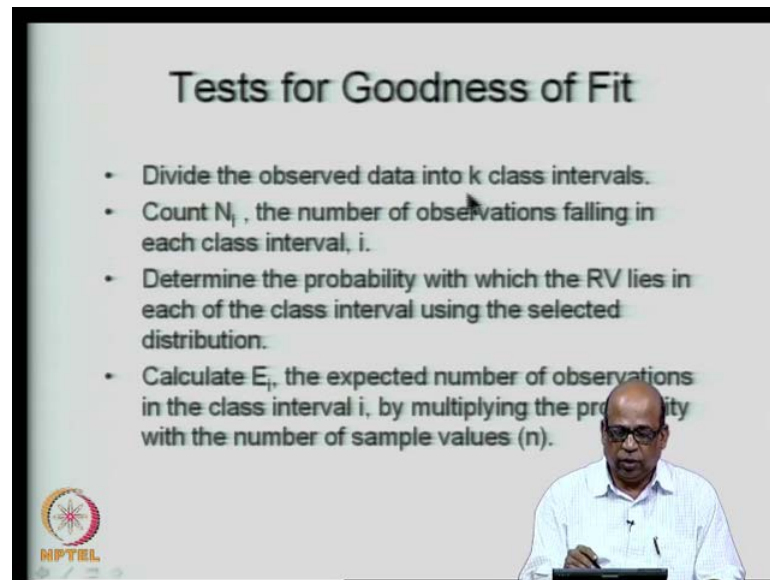
How do we do this? Let us say you have 50 years of monthly data, monthly stream flow values, which means, six hundred values you have. The six hundreds values, that we have, we divided it into a number of class intervals. Let us say my total range of values is between zero to one thousand, so six hundreds value are all within this range zero to one thousand. That means, one thousand is a highest magnitude that you have in the observed data. We may divide this into, let us say, ten class intervals, 0 to 100, 100 to 200, 200 to 300 and so on. So, like this, class, ten class intervals you may divide and then look at how many values have appeared in each of the class intervals. So, that is a way we compute the relative frequency.

So, the frequency of occurrence of the values in the range 0 to 100, let us say, five number of times, 0 to 100 (()), like this corresponding to each of the class intervals. We count from the data; we count the number of values that fall into each of the class intervals.

So, the steps we will just enumerate now. First is you have to decide whether you want to make this test for normal distribution, log normal distribution, exponential distribution, gumbel distribution and so on.

So, first you decide the distribution for which you want to examine whether the data fits at particular distribution or not. Then, each distribution, there is a theoretical distribution has its own parameter and we, we have seen how to estimate the parameter based on the observed data, that you have sample estimates of the parameter.

(Refer Slide Time: 14:02)



The slide is titled "Tests for Goodness of Fit" and contains the following bullet points:

- Divide the observed data into k class intervals.
- Count N_i , the number of observations falling in each class interval, i .
- Determine the probability with which the RV lies in each of the class interval using the selected distribution.
- Calculate E_i , the expected number of observations in the class interval i , by multiplying the probability with the number of sample values (n).

The slide also features the NPTEL logo in the bottom left corner and a photograph of a presenter in the bottom right corner.

So, let us say you want to examine for normal distribution, you have two parameters, namely mean and standard deviation. So, you get the sample estimates of these parameters from the observed data and then the observed data, that you have, you divide it into some number of class intervals, let us say we call it as k number of class intervals, it may be 10, it may be 15, it may be 5, 6 and so on.

So, we divide it into k number of class intervals. Some guidelines of dividing, of dividing this into k class interval I will explain it, presently corresponding to each of the class interval. Now, we count the number n_i , which falls into a particular class interval i , as I said, 0 to 100 is one class interval. From the observed data you see how many times it went into this particular class interval. 500 to 600 is another class interval, you again examine how many times it went into class, such things.

Now, when you do the, class, classification uniformly, it is possible, that you may get no values in a particular class interval. That means, that particular class may not have been represented in the data at all, which means, you may get a value of 0 there. We will see some general guidelines of how to make the classification, so that the test can be usefully

made. Then, we will determine the probability with which the random variable lies in each of the class interval using the selected distribution.

Let us say, you are talking about the class interval 100 to 200 and you want to examine the expected relative frequency of a value belonging to that particular class interval when it is drawn from a normal distribution, for example, normal distribution with a mean of μ and standard deviation, of s , standard deviation of σ . So, we know how to compute the probability of random variables, which follows a normal distribution to take value between these two values, these two extreme values of that interval; that means 100 to 200. So, we compute that from the theoretical distribution.

Then, once we compute the probability, remember probability is what? It is a relative frequency n_i by n . So, from this we can compute the expected number of observations, that you, that the distribution leads to for that particular interval. So, this we can get from the theoretical probability that we got from that particular distribution.

How do we get, let us say, the probability of belonging to that, that particular interval? 100 to 200 is given as 0.04 or some such thing and there are numbers of values of 100. So, 0.04 into 100, which may come to 4, so like that we compute the expected number of observation. So, there is an observed number of observations in each of the class interval. Now, we also obtain, that the expected number of observation in each of the class intervals.

(Refer Slide Time: 17:20)

The slide is titled "Tests for Goodness of Fit". It contains the following text and formulas:

- The Chi-square test statistic is calculated by

$$\chi^2_{data} = \sum_{i=1}^k \frac{(N_i - E_i)^2}{E_i}$$

- This statistic follows Chi-square distribution with number of degrees of freedom equal to $k-p-1$, where p is no. of parameters of the distribution.
- The hypothesis that the data follows a specified distribution is accepted if

$$\chi^2_{data} < \chi^2_{1-\alpha, k-p-1}$$

The slide also features the NPTEL logo in the bottom left corner and a small inset image of a man in a white shirt in the bottom right corner.

Then, we compute a statistic, which is called as a Chi-square test statistic from the data. So, I indicate this as Chi-square data as the number of observation in each of the class interval minus the expected number of observation in each of the class interval the whole square divided by the expected number of observation in the class interval, like for each of the summed over class interval i is equal to 1 to k , there are k number of classes, so summed this over or the class intervals.

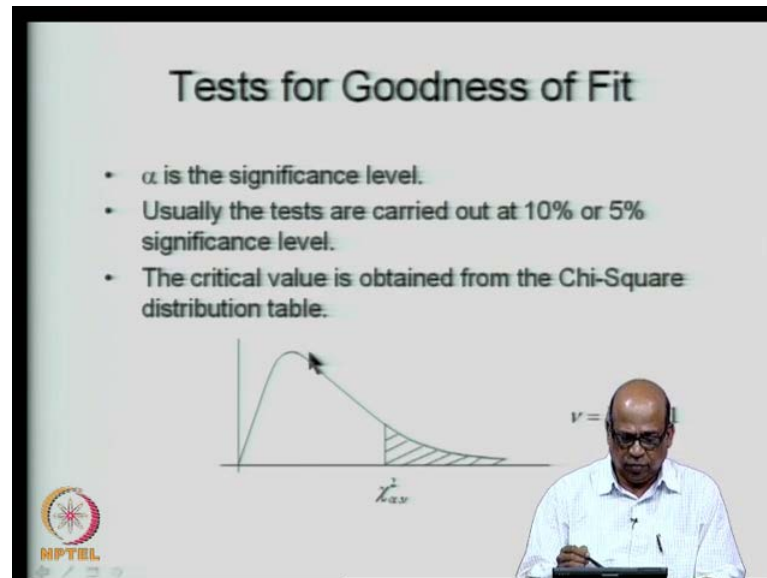
So, this gives you the Chi-square data, this subscript data indicates, that this is computed from the data. This is statistic, follows a Chi-square distribution, which has as its parameters, as a number of degree of freedom, that is, degrees of freedom of k minus p minus 1, where k is the number of class intervals and p is the number of parameters that your distribution has. Say for example, normal distribution, it has two parameters; exponential distribution, it has one parameter and so on.

So, the distribution for which you are testing the number of parameters of that particular distribution is p . So, k minus p minus 1 is the degrees of freedom and we do this test for a certain significance level, which means we are saying, that we want to test whether the data follows a normal distribution at 10 percent significance level, at 5 percent significance level, which in, in the sense indicates, in some rough sense indicates 90 percent confidence, 95 percent confidence and so on. So, we are able to say, that this particular data follows a normal distribution with that specified significance level of, say 10 percent.

Now, there is a critical level of the Chi-square value, which is associated with the degree of freedom, which is k minus p minus 1, k is the number of class interval of your data, p is the number of parameters of the distribution and α is a significance level, that you are trying making the test for. So, α can be 0.1, if it is the 10 percent significance, α can be 0.1. If it is 20 percent significance, α is 0.2 and so on. So, typically, you use 10 percent and 5 percent significance levels. Now, these are the critical values. So, you obtain the Chi-square value from the data and then examine with respect to the critical value. If the obtained Chi-square value from the data is less than the critical value, then you accept the hypothesis.

What is the hypothesis? That, the data comes from a distribution, data comes from a population, which follows a given distribution for which you are making the test. So, we accept the hypothesis that the data comes, in fact, from that particular distribution.

(Refer Slide Time: 20:40)



The slide is titled "Tests for Goodness of Fit". It contains three bullet points:

- α is the significance level.
- Usually the tests are carried out at 10% or 5% significance level.
- The critical value is obtained from the Chi-Square distribution table.

Below the text is a graph of a Chi-square distribution curve. The curve starts at the origin, rises to a peak, and then tapers off to the right. A vertical line is drawn at a point labeled $\chi^2_{\alpha, \nu}$ on the x-axis. The area under the curve to the right of this line is shaded with diagonal lines, representing the rejection region. An arrow points to the peak of the curve. To the right of the graph, the text $\nu =$ is visible. In the bottom left corner of the slide, there is a logo for NPTEL. In the bottom right corner, a man in a white shirt and glasses is shown from the chest up, looking at a tablet device.

See, the Chi-square distribution looks something like this and this is the Chi-square alpha nu, where nu is the degree of freedom k minus p minus 1. So, we would like to have our computed Chi-square in this region.


So, any value of Chi-square less than this, you accept the test and any value of Chi-square greater than this, you reject the hypothesis for a given alpha value, which is a significance level, and for given mu value, which is the degrees of freedom. So, as I mentioned, it is usually carried out at 10 percent significance or 5 percent significance level.

The tables for these are available in any standard text books on statistics. So, we use the table and from the computed value we just compare it with the critical values of Chi-square that are available from the table. The table gives specifically Chi-square values for a given alpha and nu value, which we know how to compute.

(Refer Slide Time: 21:50)

Tests for Goodness of Fit

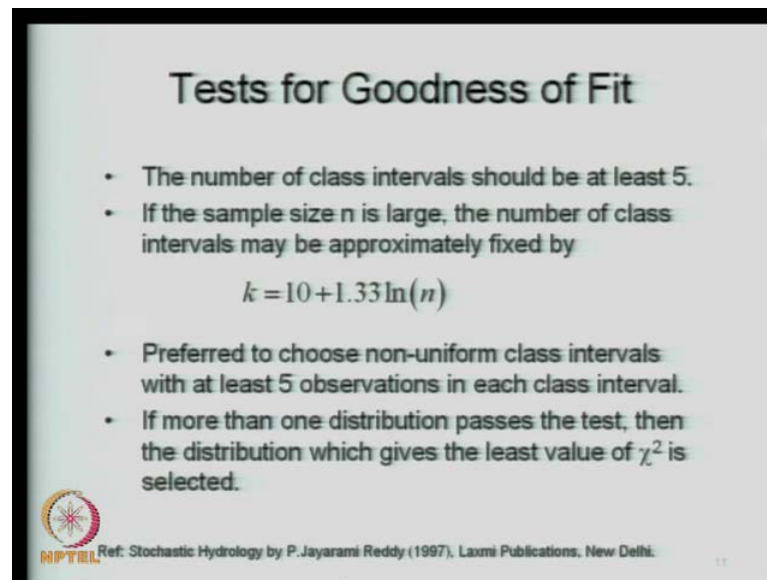
$\nu \backslash \alpha$	0.995	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	3.9E-05	0.004	0.016	2.71	3.84	5.02	6.63	7.88
2	0.010	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.21	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	3.94	4.87	15.99	18.31	20.48	23.21	25.19
20	7.43	10.85	12.44	28.41	31.41	34.17	37.57	40.00
30	13.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67



Now, this is how the tables look, say alpha is 0.995, 0.95, etcetera. Typically, we will be using 0.1, 0.5, etcetera. So, this is 10 percent, this is 5 percent significance, 2.5 percent significance and so on. And on the other side you have nu, which remember, depends on the number of class interval that you have. Let us say nu is k minus p minus 1, k is the number of class intervals.


Let us say k is 10, 10 number of class interval, then p is the number of parameter, 2, so 10 minus 2 minus 1, 3. So, you go to nu is equal to 3 and associated with that particular significance level you compute, you get the critical value of Chi-square. So, these values are typically available in most text books, in standard text books of statistics, the number of class intervals.

(Refer Slide Time: 22:44)



Tests for Goodness of Fit

- The number of class intervals should be at least 5.
- If the sample size n is large, the number of class intervals may be approximately fixed by
$$k = 10 + 1.33 \ln(n)$$
- Preferred to choose non-uniform class intervals with at least 5 observations in each class interval.
- If more than one distribution passes the test, then the distribution which gives the least value of χ^2 is selected.

 NPTEL Ref: Stochastic Hydrology by P. Jayarami Reddy (1997), Laxmi Publications, New Delhi.

Now, we will come to the basic question of how do we divide it into some number of class intervals, which is good for the test, that you are carrying out. You must remember, that the test, that we carry for the observed data will be quite sensitive to the distributions of values in these class intervals.

So, there are some guidelines available, specifically if you are interested in the central region of the distribution. Let us say it may be normal distribution or log normal distribution, etcetera, you are not really interested in the extreme values, but you are interested mostly in the central region of the values. Then, you make sure, that your central region has a fair representation in the type of class interval division that you do.

The central values are more or less uniformly distributed. In fact, one rule of thumb, that in most of the hydrologic application, that we use is that each of the class intervals, that we have, must have at least five values, which means, you divide it, you divide your data into so many class intervals, that each of the class intervals has at least five values, which also means, that the class intervals need not be uniform, like I said, 0 to 100, 100 to 200, 200 to 300. This is the mechanical way of doing it, simply divided into so many class intervals, uniform, of uniform length or uniform intervals.

But when you do that there may be some class intervals, which may have 0 or 1 values, whereas other class intervals may have 20, 30 values and so on. So, you must do this rather iteratively and then satisfy yourself, that each of the class intervals, at least, have

five values, which means you may have to combine two or three class intervals, you may have to combine and then bring it to a given acceptable interval.

So, the first guideline is that the class interval we are using need not be uniform, they can, they can be and in fact, in most of the application they are non-uniform. And each of the class interval, a general broad guideline is that it should have at least five values, which means, let us say, you have 40 years of observed data, 40 years of annual stream flow data, which means, forty values you have. Let us say you divide it mechanically into class intervals, then you may expect, that each of the class interval should have at least five class intervals if the observed data, in fact, is uniformly distributed or across the intervals.

But it may so happen that some of the class intervals may have zero values, some of them may have one value and so on in which case you combine these class intervals, and rather than having eight class intervals, you may just have five class intervals, non-uniform class intervals.

So, like this you may have to do this in an iterative manner or just look at the data and then divide it into some of the, one rule of thumb of how to divide it into class intervals, how many class intervals, that you want is given by $10 + 1.33 \log n$, where n is the number of observations. Now, this is just a rule of thumb, again you have to examine whether this type of division, in fact, is satisfactory in terms of your obtaining at least five observations in each of the class interval.

Then, when we do this test, let us say, that we are doing it for normal distribution first and we carry out the test, how to carry out step-wise I will explain it presently, but we carry put the test and then, let us say, it follows, it passes the test that it follows normal distribution. That is the hypothesis that the data comes from the normal distribution, is accepted when we carry out this test. Let us say we progress further and examine the data for log normal distribution. Let us say, it also passes the test of log normal distribution, then gamma distribution, it also passes for gamma distribution, exponential distribution, it also passes for exponential distribution, like this if you have the test passing or the acceptability or hypothesis being true for several distributions, one distribution, two distribution, three distribution and so on, then which one do we choose?

Look at this critical value of Chi-square data. Now, this, in some case indicates the acceptability of that particular statistic as obtained from this. So, you accept the particular distribution, you use the particular distribution, which gives the least value of Chi-square, that is, one general guideline. So, first you divide the data into some number of class intervals, make sure that each of the class intervals has reasonable number, which is five numbers of observations in this particular case and then make sure also, that your sample size is reasonably large.


That means, as I mentioned, if you had forty values and divided into eight class intervals, you may expect if they are all informally distributed across the region, then you may expect at least five values. Let us say, that you have only twenty values and then you divided into eight class intervals, then there may be many class intervals, which are not represented at all, in which case you have to combine them and bring it down to five class, three class intervals and so on. For this test to be useful or for this test to be more reliable, the available data must be large in length. That means, you must have significantly large length of data. If you use this test for just ten values, twenty values and so on, then you are confidence in, in the results can be quite small and therefore, any statistical test of this type, where we are looking at goodness of fit of distributions, must be essentially done on large length of data.

(Refer Slide Time: 29:24)

Example – 1

Consider the annual maximum discharge x (in cumec) of a river for 40 years. Check whether the data follows a normal distribution using Chi-Square goodness of fit test at 10% significance level.

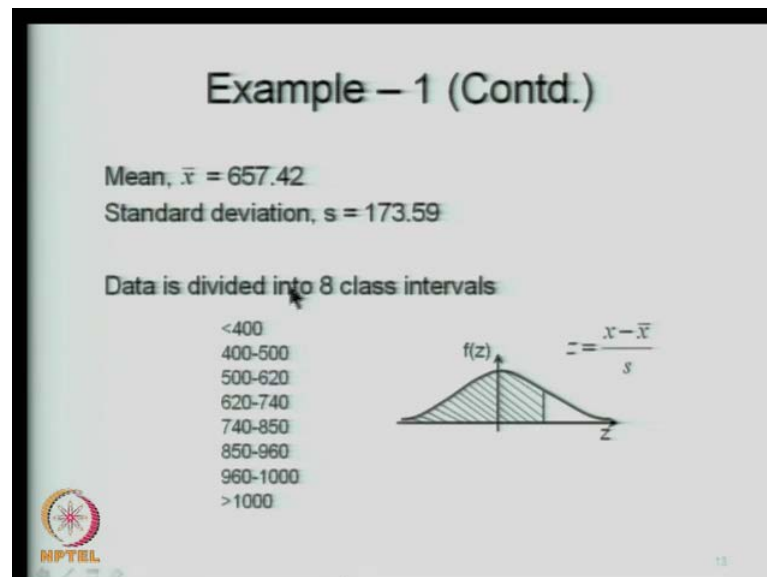
S.No.	x ($\times 10^6$)	S.No.	x ($\times 10^6$)	S.No.	x ($\times 10^6$)	S.No.	x ($\times 10^6$)
1	590	11	501	21	863	31	658
2	618	12	360	22	672	32	646
3	739	13	535	23	1054	33	1000
4	763	14	644	24	858	34	653
5	733	15	700	25	285	35	626
6	318	16	607	26	643	36	543
7	791	17	686	27	479	37	650
8	582	18	411	28	613	38	900
9	529	19	556	29	584	39	765
10	895	20	512	30	900	40	831



So, let us see how we use the Chi-square test for fitting the distribution with an example and we will take the normal distribution, we are looking at the annual maximum discharge x for 40 years, so this is the 40 years data. So, this is the serial number, and this is the value, it is in cubic meters per second, so into 10 to the power 6 cubic meter per second. So, there are forty values now.

We want to examine whether this sample of observed data comes from a population that fits normal distribution, that comes from a normal distribution and we will examine this using the Chi-square test at 10 percent significance level.

(Refer Slide Time: 30:16)



So, as I said, first we divide it into a number of class intervals, let us say, we arbitrarily choose eight class intervals in this particular case. So, there are forty values and we choose it as eight class intervals. Look at the actual range of values and then, using your judgment we will say, that less than 400, we call it as class number 1, 400 to 500, 500 to 620, 620 to 740, etcetera. As you can see, they are not uniformly; they do not have same length.

We also compute, because we are talking about normal distribution, we compute the parameters of the normal distribution or we estimate the parameters of the normal distribution from the sample. So, these are the sample estimates. \bar{x} is 657.42 and standard deviation s is 173.59. Remember, the basis is that we want to examine whether the observed relative frequency or the observed number in each of these class intervals


closely approximates the expected number in these, in the corresponding class intervals. If it follows normal distribution, that is a basis now, observed relative frequencies, we can directly obtain from the data. For example, I know how many of those forty values are less than 400, how many of them fall in 400 to 500 and so on. So, that is directly from the data.

Now, from the theoretical distribution we must be able to get what is the expected relative frequency of these particular class intervals. That means, if we randomly select a value from this particular distribution having these parameters, then we must know how many times, that random variable can take values between 400 to 500; how many times it can take from 500 to 620 and so on. So, corresponding to each of the class intervals we have on the one hand observed relative frequency and on the other, we have the expected relative frequency or in this particular case we directly have the observed number and the expected number.

(Refer Slide Time: 32:34)

Example – 1 (Contd.)

Class Interval	N_i	(z_i) of upper limit	$F(z_i)$	$p_i = F(z_i) - F(z_{i-1})$	$E_i = np_i$	$\frac{(N_i - E_i)^2}{E_i}$
<400	3	$\frac{400-657.42}{173.59} = -1.483$	0.069	0.069	2.76	0.021
400-500	2	-0.907	0.182	0.113	4.52	1.405
500-620	12	-0.216	0.415	0.233	9.32	0.771
620-740	12	0.476	0.683	0.268	10.72	0.153
740-850	4	1.109	0.866	0.183	7.32	0.06
850-960	5	1.743	0.959	0.093	3.72	0.440
960-1000	1	1.974	0.976	0.017	0.68	0.54
>1000	1		1	0.02	0.8	0.04
Total	$n = 40$			1		



So, how do we get the expected number expected number is simply, as I said, probability of that particular interval being represented multiplied by the total number because the probability is simply relative frequency n_i by n . So, n_i can be got as n_i by n , which is the probability multiplied by n , which is the total number of observation. So, from the theoretically probability, you multiplied by the total number of observation to get the expected number. So, this is how we do the examples.

Now, we have the class intervals, eight class intervals, less than 400, 400 to 500 and so on; we have eight class intervals. The n_i is the observed relative, observed numbers, this is from the observed data. So, what I mean by this is that you have three numbers of values, which are less than 400. Similarly, two less than lying between 400 to 500 and so on, like this we do the observed, data observed number.

Then, we compute the Z corresponding to the upper value of this, that we call it as F of Z i . How do we compute Z? This is simply our x minus \bar{x} by s . So, we use the higher value, 400 in this case, minus \bar{x} 657.42 divided by s , which is 173.59. So, you get a Z of minus 1.483, so this is from the upper limit. What does this give? This gives F of Z, capital F of Z, which is, from this you can get capital F of Z, that is, from the Z you can get capital F of Z up to this point, that is, probability of Z being less than or equal to the given value of Z, that is what it gives from the F of Z.

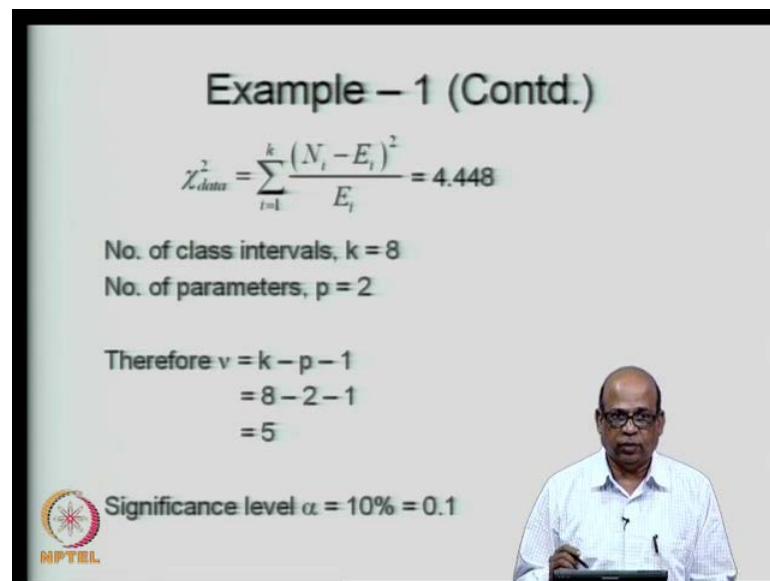
You can calculate the probability of that particular interval, for the first value the F of Z itself gives the probability, that means, you are starting from F of Z is equal to 0, you are going up to this point, 0.069. So, this gives the probability of the value of class interval falling in this particular interval, that is, I will repeat, that the probability, that a value randomly chosen will fall in the class interval 0 to 400 is 0.069, that is what you get here. Then, from this you get the expected number.

So, this is, from the observed you got the observed number and this is the expected number. So, the expected number, as I said, we will, you will get it based on the probability multiplied by the number. So, n into p_i , so n is 40 here, n is a number of observation into p_i , and then we will get n_i , which is the observed data minus E_i whole square divide by i .

But remember, here let us say, you go to class interval two. Now, class interval two, you have 0.182 associated with this Z value, how do we get this? First, you get the Z values corresponding to the higher limit of this class interval 500 and then, that is minus 0.907, associated with minus 0.907 you get F of Z i , which is 0.182. Then, F of Z i minus F of Z $i-1$, so my 0.182 minus 0.069, that is what you will get as 0.113. So, you use the previous two values and then you get 0.113. Once you get 0.113, then you multiply that by your n value, sorry for this, and then you get expected number of values.

Remember, here the expected number need not be an integer because we are taking it out from a theoretical distribution. So, you will get numbers such as 2.764, 4.52 and so on. So, although we call it as a number, it may not be an integer here, it is just an expected number of values. Now, we calculate n_i minus E_i the whole square divided by n_i , which means, this particular case, you have 3 minus 2.76 the whole square divided by 2.76, which is 0.021, like this you calculate all of these, the sum of that will be 4.48448.

(Refer Slide Time: 37:47)



Example - 1 (Contd.)

$$\chi^2_{data} = \sum_{i=1}^k \frac{(N_i - E_i)^2}{E_i} = 4.448$$

No. of class intervals, $k = 8$
No. of parameters, $p = 2$

Therefore $v = k - p - 1$
 $= 8 - 2 - 1$
 $= 5$

Significance level $\alpha = 10\% = 0.1$

NPTEL

This is the Chi-square value that you obtain from the data. Now, this Chi-square value, we will, so Chi-square from the data indicated as Chi-square data is 4.48, which is just the sum of these values.

Now, the number of class intervals we have decided as eight and the number of parameters is two because we are talking about normal distribution, mean and standard deviation. And therefore the degrees of freedom which is ν is k minus p minus 1 which is 8 minus 2 minus 1 which is five and we are testing it at a significance level of 10 percent which is α is point 1. So, we have degree of freedom of 5 and the significance level of point 1 we enter the Chi-square table and then, obtain the critical value of the Chi square.

(Refer Slide Time: 38:36)

Example – 1 (Contd.)



From the Chi-square distribution table,

$$\chi_{0.1,5}^2 = 9.24$$

$\nu \backslash \alpha$	0.9	0.1	0.05
3	0.584	6.25	7.81
4	1.06	7.78	9.49
5	1.61	9.24	11.07
6	2.20	10.64	12.59

$$\chi_{data}^2 < \chi_{0.1,5}^2$$

The hypothesis that the normal distribution can be accepted at 10% significance level



So, this part of the Chi-square table is reproduced here. These tables, I repeat, are available in any standard text books. So, you must only know how to read the tables. So, for reading the table you need the degrees of freedom ν , which is calculated as k minus p minus 1 and the significance level α .

So, we are doing the test for the significance level of 0.1 as the degrees of freedom as we obtain is 5. So, associated with 5 of ν and 0.1 of α , you get the critical value of Chi-square as 9.24 from the table. So, this is Chi-square, 0.15 comes to be 9.24 and what is our Chi-square data? It is 4.448. So, the Chi-square data 4.448 is less than the critical Chi-square value, which is 9.24 and therefore, we accept the hypothesis, that this sample, in fact, comes from a population having normal distribution. So, that is the way we carry out the Chi-square test.

We demonstrated the Chi-square test with normal distribution. Let us say, you wanted to do it for exponential distribution, what is the difference? The difference lies in only the expected relative frequency, how we compute the expected relative frequency. So, we must know how to compute the expected relative frequency associated with any of the probability distribution, which we, from our earlier basics of probability distribution we know. For example, exponential distribution, you know, capital, **f of z**, F of x , which is the c d f of that and from the c d f you get the probability of x being less than equal to x . And therefore, between two class intervals, let us say, you are talking about 100 and 200,

so you know the probability, that it lies in that particular interval of it, in fact, came from exponential distribution.

So, the only difference between distribution to distribution lies in the way we compute the F of Z. Here, in this particular case, we call it as Z, we can also call it as x if it, there is any other distribution we call it as Z here because we are related to the normal distribution. So, given any distribution we must know how to compute the expected number, that can be, that comes into that particular interval and we know how to compute these, because once we know the probability, you know the total number of values and therefore, you can calculate the expected number of values.

So, this test you can do it for any of the distributions, like normal gamma distribution, exponential distribution, etcetera all the distribution. We know how to compute the capital F of Z, maybe from analytical method or maybe from numerical integration methods and so on, the tables for which are available. So, we know how to compute the capital F of z, F of x, which is the c d f value. From the c d f value we can compute the expected number of value, that fall in a particular interval and then we compute the associated Chi-square value, compare the Chi-square value, that you compute from the data with the critical Chi-square value, which depends on the significance level at which you are doing the test as well as on the degrees of freedom given by k minus p minus 1, apart from the expected number of values, that change from distribution to distribution.

The p, which is the number of parameters, also changes from distribution to distribution, exponential distribution, for example, has only one parameter lambda. So, p will be 1 log, normal distribution will have two parameters; gamma distribution will have three parameters and so on. So, we know now how to carry out the test.

(Refer Slide Time: 42:40)

The slide is titled "Tests for Goodness of Fit". It describes the Kolmogorov-Smirnov Goodness of fit test. The text on the slide is as follows:

Kolmogorov – Smirnov Goodness of fit test:

- Alternative to the Chi-square test.
- The test is conducted as follows:
 - The data is arranged in descending order of magnitude.
 - The cumulative probability $P(x_i)$ for each of the observations is calculated using the Weibull's formula.
 - The theoretical cumulative probability $F(x_i)$ for each of the observation is obtained using the assumed distribution.

In the bottom left corner, there is a logo for NPTEL (National Programme on Technology Enhanced Learning) featuring a stylized sun or starburst design. In the bottom right corner, the number "17" is visible.

There is another test, that we will discuss now, which is called as the Kolmogorov-Smirnov test. This is an alternative to the Chi-square test. Normally, when we want to examine whether the data fits a given distribution or not, we carry out both the tests for reasons I will come to slightly later. Kolmogorov-Smirnov test is advantageous computationally and also, in many situations it is a superior test statistically.

So, the, in Chi-Square test, what we did is that we compared the relative frequencies with the observed frequency and then obtained the statistic out of it. What we do here is that again, we, rather than dividing into class intervals we use the data as they are observed and then compute the probabilities of those values occurring.

So, here, like we did in the probability plotting position methodology, what we do is we arrange the data in a descending order, highest value first and the lowest the last. Then, the cumulative probability associated with each of these is calculated from the Weibull's formula, that is, what is the Weibull's formula?

It is the small p is equal to m divided by n plus 1, this is what we discussed in the last class. So, small p is equal to m divided m plus 1, what does it give? It gives probability of x being greater than equal to x , that is the particular value of x , but we want the cumulative probability, that is, F of x is what we need, that is, x being less than equal to X and therefore, we take 1 minus p of that and the get the cumulative probability associated with that particular value of x i.

So, the first difference that we see from Chi-square, between Chi-square and Kolmogorov is in the Kolmogorov test. We are not doing classification, we use the values as they are and arrange them in the descending order, compute associated with each of the values, the f of x , capital F of x , which is the c d f value and this c d f value we compute based on the Weibull's formula. This gives you the observed cumulative probability, but we also have a theoretical cumulative probability associated with each of the observation. This we get it from the parent distribution, the distribution for which we want to make the test. For example, for the normal distribution we know how to compute the cumulative distribution function associated with that particular value. So, we calculate the cumulative theoretical cumulative probability.

(Refer Slide Time: 45:32)

Tests for Goodness of Fit

- The absolute difference of $P(x_i)$ and $F(x_i)$ is calculated.
- The Kolmogorov-Smirnov test statistic Δ is the maximum of this absolute difference.

$$\Delta = \text{maximum} |P(x_i) - F(x_i)|$$

- The critical value of Kolmogorov-Smirnov statistic: Δ_0 is obtained from the table for a given significance level α .
- If $\Delta < \Delta_0$, accept the hypothesis that assumed distribution is a good fit at level α .

And then, we take the absolute difference between the observed cumulative probability and the expected cumulative probability. So, this is from the observed data. Using the Weibull's formula we get P of x_i and this is from the theoretical distribution F of x_i and we take the absolute difference and the test of statistic for the Kolmogorov test is denoted as delta.

We take the maximum absolute difference between P of x_i and F of x_i and this maximum difference, absolute difference, we compare it with critical value of the KS statistic, Kolmogorov-Smirnov statistic is also abbreviated as KS statistic. So, this KS statistic delta naught, again much like Chi-square test, Chi-square statistics, they are

available in standard text books. So, we can pick corresponding to a given significance level of alpha. So, delta naught just depends on significance level of alpha.

Now, if the delta, that you have calculated thus, as an absolute, as a maximum absolute difference between the observed c d f and the expected c d f value associated with the value x I, if delta is less than delta naught, then we accept the hypothesis, that the assumed distribution is a good fit at significance level alpha.

Remember, when we are making conclusions, always you, you must remember, that we are doing it at a particular significance level. Let us say the test fails at 5 percent level, but it passes at 10 percent significance level, so it is possible, that then you may have a 90 percent confidence and now 95 percent confidence, that it fits that particular distribution. So, this you must always keep in mind, these tests are associated with a certain degree of confidence that you have and therefore, they are carried out at that particular significance level Now, therefore, they are, their confidence level is associated with the significance level at which you are carrying out the test.

(Refer Slide Time: 48:06)

Tests for Goodness of Fit

Table for Kolmogorov-Smirnov statistic Δ_n :

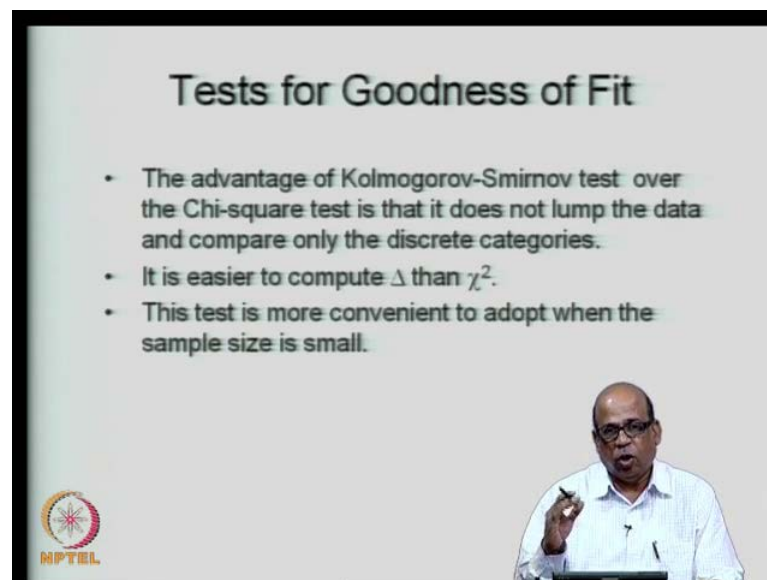
Size of sample	Significance Level α				
	0.2	0.15	0.1	0.05	0.01
5	0.45	0.47	0.51	0.56	0.67
10	0.32	0.34	0.37	0.41	0.49
20	0.23	0.25	0.26	0.29	0.36
30	0.19	0.20	0.22	0.24	0.29
40	0.17	0.18	0.19	0.21	0.25
50	0.15	0.16	0.17	0.19	0.23
Asymptotic formula ($n > 50$)	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Now, the statistic delta naught, which is the critical statistic against which you are comparing your computed delta, they are available in the tables and this is how the tables look. It is dependent on the sample size.

Remember, your Chi-square depended on the number of class interval in terms of the degrees of freedom that you had because Chi-square $k - p - 1$, that is the degrees of freedom and alpha, that is how your Chi-square tables are prepared, but the Kolmogorov-Statistic depends on the sample size itself.

Do you have forty values? Do you have hundred values, fifty values and so on? So, based on the samples size and the significance level you can get the Kolmogorov-Statistic Δ naught, these are typically given up to fifty values. And then, for more than fifty values, when you have n greater than fifty, you have some asymptotic values 1.07 divided by root n and so on. So, if you have more than fifty values, you use these KS statistic as critical values Δ naught.

(Refer Slide Time: 49:16)



The slide is titled "Tests for Goodness of Fit". It contains three bullet points:

- The advantage of Kolmogorov-Smirnov test over the Chi-square test is that it does not lump the data and compare only the discrete categories.
- It is easier to compute Δ than χ^2 .
- This test is more convenient to adopt when the sample size is small.

In the bottom right corner, there is a small video inset showing a man in a white shirt and glasses speaking. In the bottom left corner of the slide, there is a logo for NPTEL.

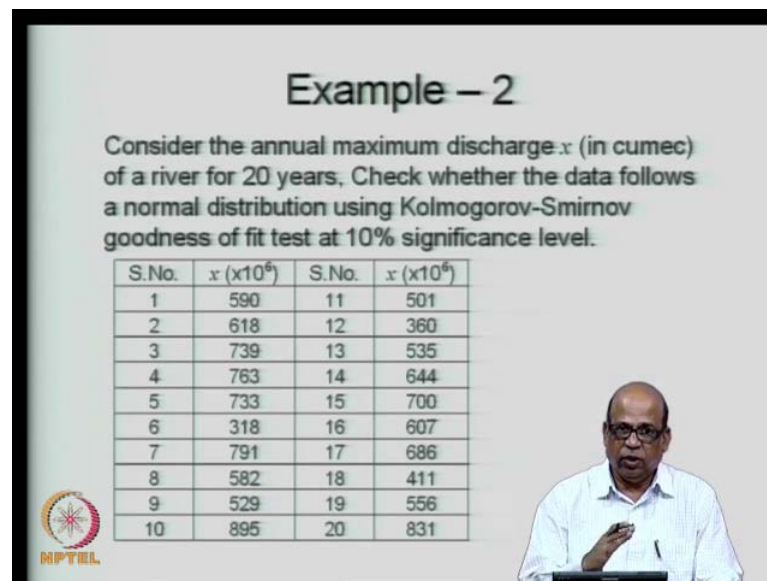
Now, the advantage of Kolmogorov-Smirnov test is, first of all, look at the competitions. All we have done is you have arranged it in decreasing order and then assigned the ranks and then calculated the observed p d f, observed F of x , which we denoted as capital P of x . And then, you also know how to get F of x from the theoretical distribution. So, all you are doing is get the distribution differences P of x minus F of x and that become one value out of that. So, doing it is very simple.

Also, you do not lump the data, that means, because your you were in Chi-square test, you were distributing it into number of class intervals, which means, anything that falls between 0 and 100. You are lumping all of data, all of that data into one class interval

and then, based on this class interval expected frequency etcetera, you are calculating. So, in the Kolmogorov-Smirnov test you do not do that, you use the observed data as they are. So, that is one advantage.

Now, when the sample size is small, the KS test is more convenient to handle compared to the Chi-square test. So, these are some of the advantages.

(Refer Slide Time: 50:44)



Example – 2

Consider the annual maximum discharge x (in cumec) of a river for 20 years. Check whether the data follows a normal distribution using Kolmogorov-Smirnov goodness of fit test at 10% significance level.

S.No.	$x (x10^6)$	S.No.	$x (x10^6)$
1	590	11	501
2	618	12	360
3	739	13	535
4	763	14	644
5	733	15	700
6	318	16	607
7	791	17	686
8	582	18	411
9	529	19	556
10	895	20	831

The slide also features the NPTEL logo in the bottom left corner and a photograph of a lecturer in the bottom right corner.

And in some statistical sense, the Kolmogorov-Smirnov test is superior to Chi-square test in many situations. That means, what I mean by that is if your Kolmogorov-Smirnov passes and the Chi-square test fails, you may go with the Kolmogorov-Smirnov hypothesis, Kolmogorov-Smirnov result, test result and vice versa. Let us say, that Chi-square test passes, but the Kolmogorov-Smirnov fails, in which case you should respect the Kolmogorov-Smirnov test more.

So, we will do the example. We will do the example now. We will take 20 years of data and then look at how we carry out the KS test. Now, these are the data values, first we arrange the data values. I will explain everything through one table. We want to carry out the KS test at 10 percent significance level to examine whether this particular data can be assumed to follow a normal distribution.

(Refer Slide Time: 51:40)

Example – 2 (Contd.)

Mean, $\bar{x} = 619.62$
 Standard deviation, $s = 153.32$

Data is arranged in the descending order and a rank (m) is assigned to each data point. The probability is obtained using Weibull's formula:

$$p = P(X \geq x_m) = \frac{m}{n+1}$$

$$z = \frac{x - \bar{x}}{s}$$

We compute the mean and the standard deviation because we need it for calculating of, calculation of F of x, capital F of x because we are following normal distribution, because it is normal distribution. We need these two parameters, then we arrange the data in decreasing order and associated with each of the value, we first calculate p, small p, which is probability of **exceedence**. Remember, x is greater than equal to x n associated with that particular rank m and then, from p we can get the c d f value, which is 1 minus p and then associated with this particular value of x we can also get the z value. And therefore, we can get the expected F of x, capital F of x.

(Refer Slide Time: 52:27)

S.No. (i)	x (x10 ³)	Descending order	Rank (m)	$p = \frac{m}{n+1}$	$P(x_i) = 1-p$	(z _i)	F(z _i)	$ P(x_i) - F(z_i) $
1	590	895	1	0.048	0.952	1.795	0.964	0.012
2	618	831	2	0.095	0.905	1.381	0.916	0.011
3	739	791	3	0.143	0.857	1.121	0.869	0.012
4	763	763	4	0.190	0.810	0.936	0.825	0.015
5	733	739	5	0.238	0.762	0.781	0.783	0.021
6	318	733	6	0.286	0.714	0.737	0.769	0.055
7	791	700	7	0.333	0.667	0.524	0.700	0.033
8	582	686	8	0.381	0.619	0.432	0.667	0.048
9	529	644	9	0.429	0.571	0.162	0.564	0.007
10	895	618	10	0.476	0.524	-0.011	0.495	0.029
11	501	607	11	0.524	0.476	-0.082	0.468	0.008
12	360	590	12	0.571	0.429	-0.192	0.424	0.005
13	535	582	13	0.619	0.381	-0.242	0.404	0.023
14	644	556	14	0.667	0.333	-0.412	0.340	0.007
15	700	535	15	0.714	0.286	-0.549	0.292	0.006
16	607	529	16	0.762	0.238	-0.589	0.278	0.040
17	686	501	17	0.810	0.190	-0.772	0.220	0.030
18	411	411	18	0.857	0.143	-1.361	0.087	0.056
19	556	360	19	0.905	0.095	-1.692	0.045	0.050
20	831	318	20	0.952	0.048	-1.965	0.025	0.023

So, that is what we do. Now, this is i , i is equal to 1 to 20, there are twenty values. These are the x values, which are the actual available data series. We arrange this data in descending order, so this is the descending order series, 895 highest, 831 and so on. We assign ranks 1, 2, 3, 4, etcetera; highest value gets the rank number 1, the lowest value gets the rank number 20. We calculate the Weibull probability p is equal to m divided by m plus 1, m is the rank here, n is the number of values, which is 20. So, we calculate all of these values. Then, we get the observed F of x , observed $c d f$, which we denote it as capital P of x of i , which is simply equal to 1 minus p , why, because p is the exceedance probability and the $c d f$ gives you the probability of x being less than equal to X and therefore, the $c d f$ value is 1 minus p as obtained. So, we get these values like this.

And then, we get the z_i corresponding using this equation x minus \bar{x} or s . So, for example, this is x minus \bar{x} , that we have, which is, 619, 619.62 divided by 153.32. So, that is how we get z_i . Once we get z_i we know how to calculate F of z_i , you go to the tables and get the F of z_i . So, you got the expected probability and you have the observed probability, you take the absolute difference between these two, 0.952 minus 0.964, which will be negative, but they are taking absolute values. So, these are the absolute differences. When you get these absolute differences, look at the maximum absolute difference, this is 0.056. This defines the delta, which is the KS statistic.

(Refer Slide Time: 54:34)

Example – 2 (Contd.)

Maximum value $\Delta = 0.056$


From the Kolmogorov-Smirnov table,

$\Delta_0 = 0.26$

$N \backslash \alpha$	0.9	0.1	0.05
10	0.34	0.37	0.41
20	0.25	0.26	0.29
30	0.20	0.22	0.24

Since $\Delta < \Delta_0$,

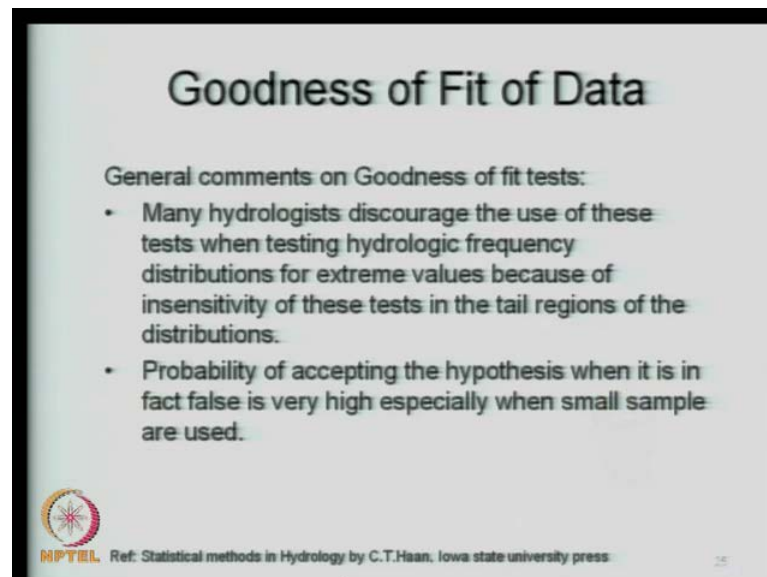
The hypothesis can be accepted that the normal distribution fits the data at 10% significance level.



So, the maximum value is Δ , which is equal to 0.056. We are doing the test at significance level of 0.1 or 10 percent significance, which is 0.1. Our sample size is 20, so we go to the Kolmogorov-Smirnov table, statistic table associated with n is equal to 20 and α is equal to 0.1. We pick up the, pick up the 0.26, so Δ_{naught} is 0.26.

The observed Δ value or Δ value coming out of the data is 0.056 and critical value is 0.26, because this is less than Δ_{naught} , we accept the hypothesis. So, the hypothesis that the data comes from a population following a normal distribution is accepted at 10 percent significance level when we follow the Kolmogorov-Smirnov test.


(Refer Slide Time: 55:39)



Goodness of Fit of Data

General comments on Goodness of fit tests:

- Many hydrologists discourage the use of these tests when testing hydrologic frequency distributions for extreme values because of insensitivity of these tests in the tail regions of the distributions.
- Probability of accepting the hypothesis when it is in fact false is very high especially when small sample are used.

 MPTEL Ref: Statistical methods in Hydrology by C.T.Haan. Iowa state university press

Now, there are two goodness of fit tests that we discussed; one is the Chi-square test and another is Kolmogorov-Smirnov test. When we are dealing with the middle range of the observations, let us say, you are talking about the monthly streams flows and then you want to plan for reservoir for water supply and so on. So, essentially you are interested in the middle range of the sequence or middle range or middle value of distribution.

These tests are quite useful, but the moment you go to the extremities, extreme values and then you want to examine, whether these extreme values for a particular distribution or not, then these tests are not really very useful. In fact, many hydrologists discourage use of these tests when we are interested in the extreme values. Also, you know, when the samples size is small, probability of accepting the hypotheses when it is, in fact, falls is quite high. So, you must be alert to the situation where your sample sizes are small and

when you are dealing with extreme values. So, these two limitations you must keep in mind, especially in applications in hydrology, before we mechanically apply these tests and then conclude, that it follow the particular distribution or not.

So, in today's class, then we examine specifically two tests, which is Chi-square test and the Kolmogorov-Smirnov test, both of which are used to examine the hypothesis, that the data, that we have comes from a population fitting a particular theoretical distribution, normal distribution, log normal distribution, etcetera.

So, the specific, that we do, we did today is to use these two tests to examine whether the data comes from a normal distribution. And these type of examples, exercises are necessary when we want to fit the data to the particular distribution and then start using all the methods of analysis, that we discussed earlier on. Because subsequently, once we assume, that it is a normal distribution we know what is to be done with the normal distribution and so on.

There are, however, limitations of using these tests. If you are interested in the extreme values we must be cautious in using these two tests. Also, the Kolmogorov-Smirnov test in many situations is a superior test compared to the Chi-square test. When the sample sizes are small, these tests should be used with caution and these tests can complement the probability paper tests that we had discussed earlier. That means, you use the data and brought it on the probability paper for that particular distribution and maybe, you also used the Chi-square test, Kolmogorov-Smirnov test, etcetera, so that your confidence in the, in the hypothesis, that this comes from the particular distribution is quite high.

So, in the next lecture we will move on slightly further away from the fitting of the distributions. We will see how we use all of this information in hydrologic designs, specifically starting with the intensity duration, frequency relationship, etcetera and then going on to (()).

Thank you very much for your attention