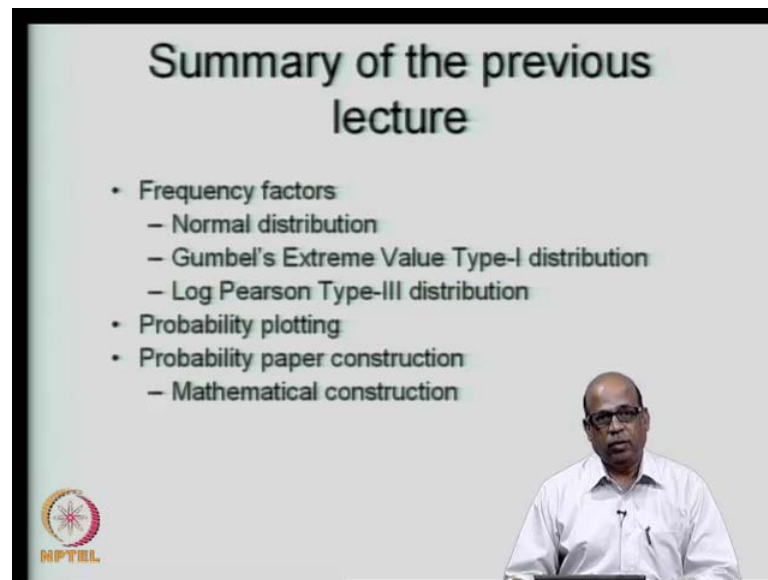


Stochastic Hydrology
Prof. P. P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

Module No. # 06
Lecture No. # 27
Probability Plotting – II

(Refer Slide Time: 00:42)



Good morning and welcome to this 27 lecture of the course Stochastic Hydrology. We have been discussing the frequency analysis and specifically, in the previous lecture, we covered the frequency factors. That is, frequency analysis by analytical methods. We discussed the frequency factors for normal distribution, for the Gumbel's extreme value type one distribution, that is e^{-v} , which is also known as a Gumbel type one Gumbel's distribution and the log Pearson type three distribution.

If you recall, for the normal distribution, the standard normal deviate z itself becomes the frequency factor. We use the frequency factors to compute the magnitudes of the flows, if you are doing the frequency analysis for the flows corresponding to any given return period.

(Refer Slide Time: 01:30)

The slide is titled "Summary of the previous lecture" and contains a bulleted list of topics. Handwritten red notes are present at the bottom right of the slide.

- Frequency factors
 - Normal distribution
 - Gumbel's Extreme Value Type-I distribution
 - Log Pearson Type-III distribution
- Probability plotting
- Probability paper construction
 - Mathematical construction

Handwritten notes in red ink:

$$p = P[x \geq x_T]$$
$$x_T = \bar{x} + K_T S$$
$$T = \frac{1}{p}$$

The slide also features the NPTEL logo in the bottom left corner and a small number '2' in the bottom right corner.

So, if you recall, we wrote the expression as x_t is equal to \bar{x} plus k_t into σ . If you are doing it for samples, we write k_t into s . Now, k_t is the frequency factor corresponding to a return period of t , capital t . So, this will depend on the return period that we are talking about. T is also equal to 1 by p . All these are necessary for carrying out the frequency analysis, where p is probability that x is greater than or equal to x_t .

Now, all of these are the preliminaries of frequency analysis. So, we are interested in getting x_t , which is the magnitude of the flow, if you are doing it for flow frequency analysis, which is typically the case. Flood frequency analysis is what we carry out. So, this is the magnitude of the flood corresponding to a return period of t . The return period of t is related to probability of exceedence by t is equal to 1 by p .

(Refer Slide Time: 01:32)

We are doing the analytical frequency analysis by using this expression \bar{x} plus k_t times s , where k_t is the frequency factor. We discussed the k_t , which is a frequency factor, corresponding to the normal distribution, which as I said is the standard normal deviate z itself. Then, for the Gumbel's extreme value type one distribution as well as log Pearson type 3 distribution, I have given you the formula for obtaining the k_t , which is the frequency factor. We also discussed subsequently the probability plotting. That is, when we have the data, how do we check whether the particular data fits a given distribution or not.

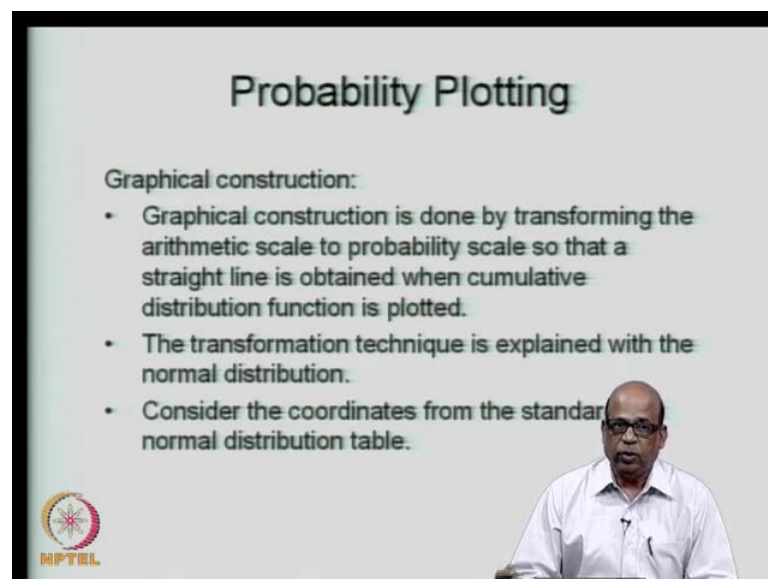
We started with the discussion on plotting the data on probability papers. Specifically, we started discussing the probability paper construction by mathematical method, that is by analytical method. If you recall, we discussed the exponential distribution to demonstrate how we construct the probability paper.

I would like to reiterate here that, essentially what we do in probability paper construction is that, we are converting the non-linear cdf into a linear form. So, the scales are adjusted on the paper, such that the cdf plots are a straight line. That is, what we did in the case of exponential distribution.

We could do that in exponential distribution because it is a kind of invertible distribution. In the sense, that given the cdf , we would be able to express x in terms of the cdf . Whereas, for many other distributions it is not readily possible for us to express x in terms of f of x given the analytical expression for f of x .

In such situations, we **afford a course** to the graphical method. What do we do in the graphical method is, let us say that you plot any cdf on an arithmetic paper. How does the cdf plot? It will typically plot as a non-linear curve. Now, this non-linear curve, we make it as a linear curve by adjusting the scales of the arithmetic paper. This adjusted scale on the arithmetic paper will convert the arithmetic paper into a probability paper.



(Refer Slide Time: 06:02)



Probability Plotting

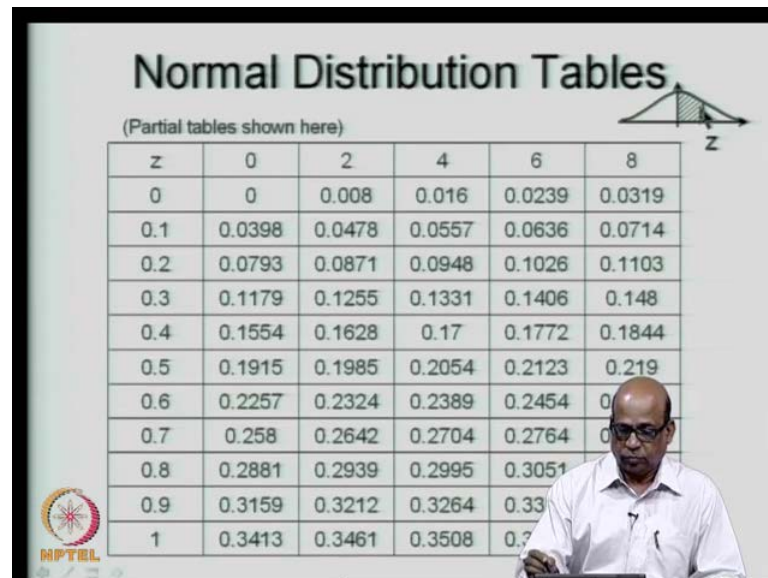
Graphical construction:

- Graphical construction is done by transforming the arithmetic scale to probability scale so that a straight line is obtained when cumulative distribution function is plotted.
- The transformation technique is explained with the normal distribution.
- Consider the coordinates from the standard normal distribution table.

We will start demonstrating this procedure by the normal distribution. So, in the graphical distribution for the normal distribution, let us say, the graphical construction is done by transforming the arithmetic scale to probability scale. So that, when you plot the c d f, you will obtain a straight line. This transformation technique, now we will discuss through the normal distribution.

(Refer Slide Time: 06:46)



Normal Distribution Tables
(Partial tables shown here)

z	0	2	4	6	8
0	0	0.008	0.016	0.0239	0.0319
0.1	0.0398	0.0478	0.0557	0.0636	0.0714
0.2	0.0793	0.0871	0.0948	0.1026	0.1103
0.3	0.1179	0.1255	0.1331	0.1406	0.148
0.4	0.1554	0.1628	0.17	0.1772	0.1844
0.5	0.1915	0.1985	0.2054	0.2123	0.219
0.6	0.2257	0.2324	0.2389	0.2454	0.2519
0.7	0.258	0.2642	0.2704	0.2764	0.2823
0.8	0.2881	0.2939	0.2995	0.3051	0.3106
0.9	0.3159	0.3212	0.3264	0.3315	0.3364
1	0.3413	0.3461	0.3508	0.3554	0.36

For the normal distribution, we have the standard normal distribution table. We make use of that and then, we look at f of z , capital f of z for various values of z . First, we plot f of z verses z , on an arithmetic paper and see how it plots. So, these are the f of z values, typically if we use a table like this.


This gives the area under the standard normal curve from 0 to that particular value of z . So, you add 0.5 to that by symmetry to get the complete area until this point. Similarly, if you are looking at a z , which is less than 0.5, you will deduct or you will use the area directly. To get the probability on the other side, for that, you will deduct this particular area from 1.

(Refer Slide Time: 07:44)

Probability Plotting

Normal distribution table:

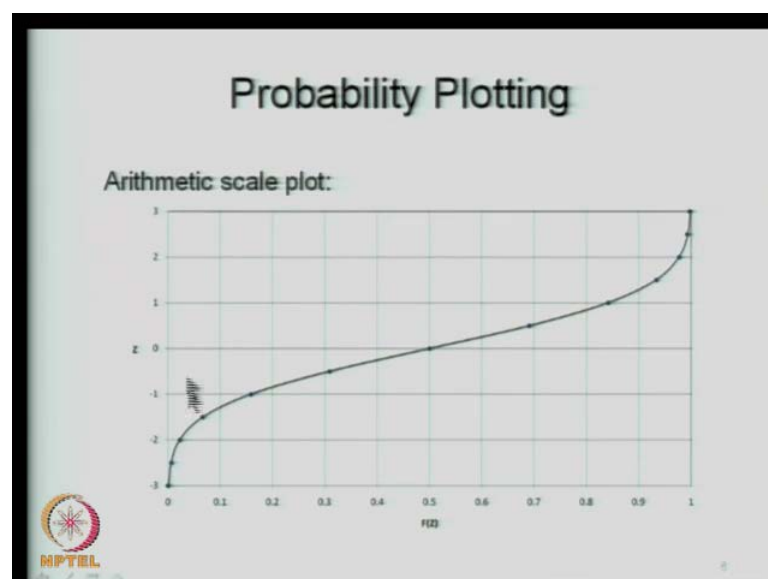
Z	F(z)	Z	F(z)
-3	0.0013	0	0.5
-2.5	0.0062	0.5	0.6915
-2	0.0227	1	0.8413
-1.5	0.0668	1.5	0.9332
-1	0.1587	2	0.9772
-0.5	0.3085	2.5	0.9938
		3	0.9987



5

So, all of these fundamentals of normal distribution we have studied earlier. So typically, this is how the table looks. These are available in any standard text books on probability and statistics. So, we have shown only a partial table here. So, for a given z , you can obtain f of z using this figure here. So, we use z and the associated z value. Let us say, from minus 3 to plus 3, about 99.9 percent of the values of the probability is contained in that.

(Refer Slide Time: 08:10)



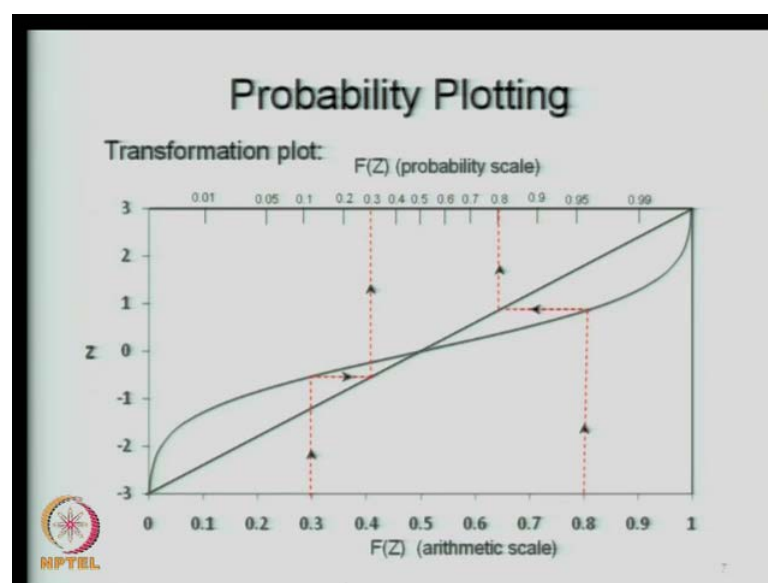
So, we use typically from minus 3 to plus 3 and these are the associated z values. This z versus f of z, we will plot first on the arithmetic paper and see how it plots. So, on the arithmetic paper, on the y axis, I plot the z and on the x axis, I plot the f of z. This is how it looks.

So, minus 3 corresponding to 0, not exactly 0, but very nearly 0 and plus 3, corresponding to nearly 1. This is a non-linear curve and at 0.5, it corresponds to 0. That is, z is equal to 0, corresponding to f of z of 0.5. Normal distribution being symmetric distribution, we have a symmetric curve plotted on the arithmetic paper.

What we will now want to do is, we want this relationship between f of z and z, to plot as a straight line. So, what is it that we do? We convert this figure. We convert the scales, in fact, on f of z such that, this figure plots as a straight line. How do we do that? We plot a straight line now.

We take a straight line passing through z is equal to 0 and f of z is equal to 0.5. We plot a straight line and then, transform the associated coordinates of f of z from this curve to that straight line. That is all. Which means, that essentially, we are transforming the f of z scale here from the arithmetic scale to another scale, which we call as probability scale, such that, this c d f, which is plotted as a non-linear curve will in fact, plot as a straight line.

(Refer Slide Time: 10:24)



So, the procedure is very simple. Once we plot the f of $c d f$, simply take a straight line which best represents that particular $c d f$. Then, **transform the coordinates** transform the scale on the f of z to a probability scale, such that, all of these points are transferred to the straight line. We will see how it is done now. Let us say, this was your plot on the arithmetic scale.

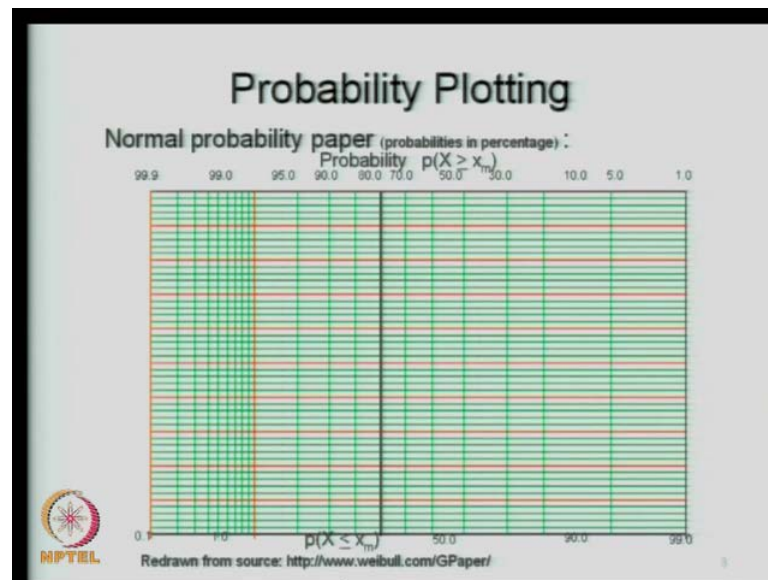
So, this curve here shows the plot, which is a same plot as is shown here. We draw a straight line between minus 3 to plus 3 here on the z axis passing through and this straight line passes through 0 and 0.5 or 0.5 f of z corresponding to a value of 0. Now, we start transforming the values on f of z .

So, f of z for example, 0.3 would have corresponded to this particular point. But we want this point to be on the straight line. So, we transfer this point to this and then, project it upwards and call this as 0.3. So, essentially what we did is that, we are saying 0.3 should not be here, but it should be here and therefore, I take it to 0.3

Similarly, 0.2. 0.2, I come and reach the curve and take it horizontally; reach this particular point. Then, say it is 0.2. Similarly, 0.1 and so on. On the other side, you take 0.8, reach the curve first and go left side and hit the straight line and project it upwards and point it.

Look at this scale now. 0.01, 0.05, 0.1, 0.2 etcetera, which are much different from the arithmetic scale. So, this is how we converted the arithmetic scale to a probabilistic scale. So, when you plot it on the probabilistic scale, the data which you have used, we will plot it as this particular straight line. Then, we prepare a probability paper now which means, the probability scale will be this and the z axis is an arithmetic scale.

(Refer Slide Time: 12:36)



So, you can use to plot any data on the arithmetic scale with the probability scale will be non-linear, as you can see. So, this is how it appears. The probability paper, normal probability paper will appear like this. These probability papers are commercially available. You can also get it from the internet and from various sources.

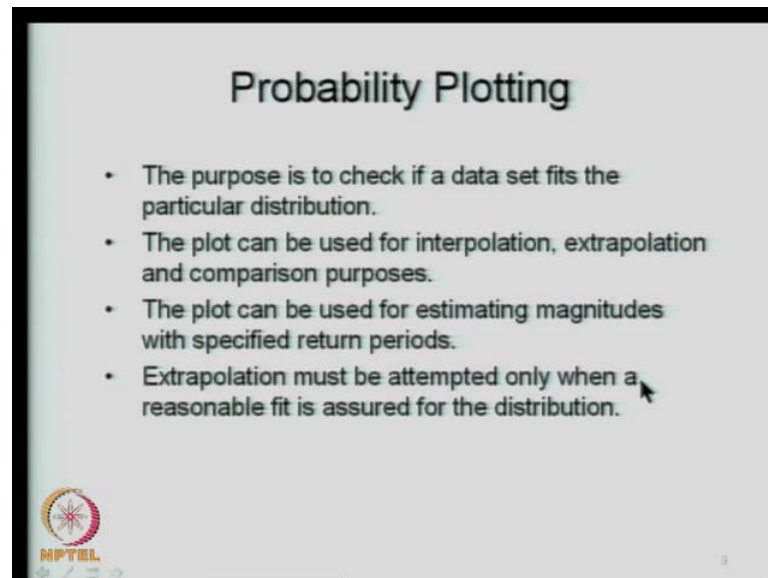
What we have done is, this one is redrawn from the Weibull.Com website, which provides the normal probability paper. As in indeed, it provides probability papers for Gumbel distribution and other distributions. Typically in such papers, the x contains the probabilities in percentages.

See in this particular probability paper, I have shown probability of x being less than equal to x m in percentage. So, 99 percent, starting with 0.1 percent and similar things are shown with 1 minus probability of x. That is, in percentage 100 minus probability of x, for probability x being greater than equal to x m. Here, it is x being less than equal to x m. So, you start with 1 and go up to 99 percent. The y axis here, as you can see, is an arithmetic scale.

So, the x axis is probability scale and y axis is arithmetic scale. We use this probability papers and plot the data on the probability paper, to check whether the given data follows normal distribution or not. What will happen if it follows the normal distribution? It would follow this particular straight line or the values will all be around this particular straight line.

So, if you plot the data that you have on the normal probability paper and it plots as a normal distribution, as a straight line on this particular probability paper, then you can reasonably assume or if it plots nearly as a straight line, then it is a reasonable assumption, that the data, in fact, comes from a population which follows normal distribution.

(Refer Slide Time: 14:46)



So, the purpose of the probability paper is to check, if a data set fits a particular distribution. Let us say you have exponential distribution probability paper, that we discussed in the last lecture and you have normal distribution probability paper. Similarly, you have typically Gumbel's distribution. In hydrology, mostly we use normal and Gumbel distribution.

So, once you have the probability paper, you fit the data on that particular paper and if the data fits as a straight line or it fits nearly as a straight line, then you can assume that the data that you have as a sample data. In fact, is drawn from that particular distribution, normal or Gumbel or whatever.

Now, once you plot this, you can use that. Let us say that, it plots as a straight line on the normal distribution. You can use that for interpolation, extrapolation and also for comparison purposes. Like, another normal distribution, how does it compare with the particular data that you have and so on. So, the probability paper is very useful especially for obtaining interpolated data.

In the sense that, let us say that, you want to talk about a flow magnitude of a 15 year return period, which from the data, it is not directly possible for it to get. Then, once you plot it as a straight line on the normal probability paper for example, you will be able to get the magnitude corresponding to 15 year flow data. That is the interpolation.

You can also use it for extrapolation. Let us say, your data range was only up to, let us say return period of 10 years, 15 years etcetera, you can obtain it straight away. But, you are interested in getting a return period, a magnitude of flow corresponding to a return period of 200 years or 300 years, which are typically the case in most hydrologic designs.

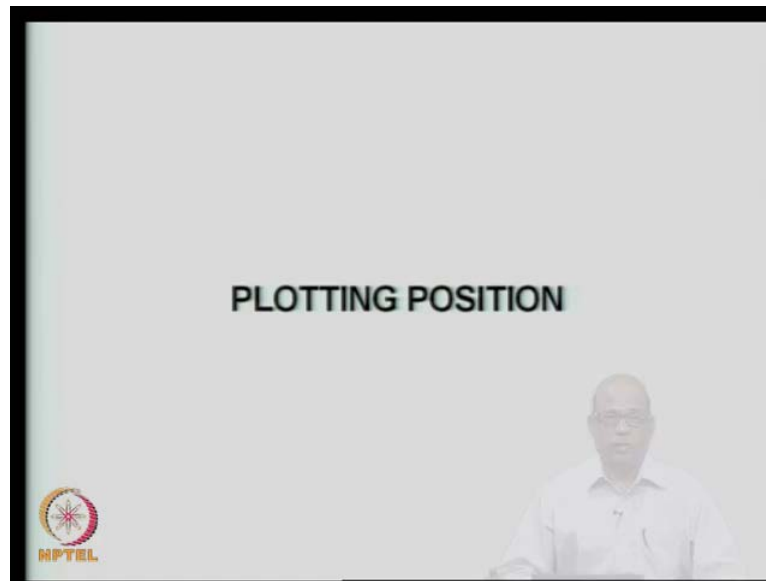
Then, you can use this procedure for extrapolation also. That is, once you have the straight line, you can also extrapolate and then, look at the return periods which are of, which far exceed what you can get from the sample data itself. That is, if you recall, what we do in frequency analysis. However, I must alert you here that extrapolation must be done very carefully.

Because, in extrapolation, essentially what we are doing is, we are looking at the extreme values. Let us say, 200 year return period or 300 year return period, etcetera. These are typically the extreme values, which are located on the tails of the distribution. Tails of the distribution are very sensitive to change in probabilities.

I will demonstrate, when I explain the example there as to what I mean by that. So, whenever you are using the probability paper for extrapolation, you must be reasonably sure that the data fits that particular distribution. If you have any doubt, especially on the extreme values departing further away from the straight line, then you have to be careful in doing the extrapolation.

Especially because, you use the extrapolation for flood frequency analysis, where the difference between a 200 year return period flow or flood and the 300 year return period flood may be significant. Therefore, the method that you use to estimate these return period floods must be accurate. Otherwise, it may lead to costly decisions in terms of the economies, economic investments and so on or it may lead to very conservative decisions, that the decisions may be wrong. Therefore, achieving the right value of the magnitude that arises out of the probability papers must be a major concern, especially, when you are dealing with the extreme values.

(Refer Slide Time: 19:14)



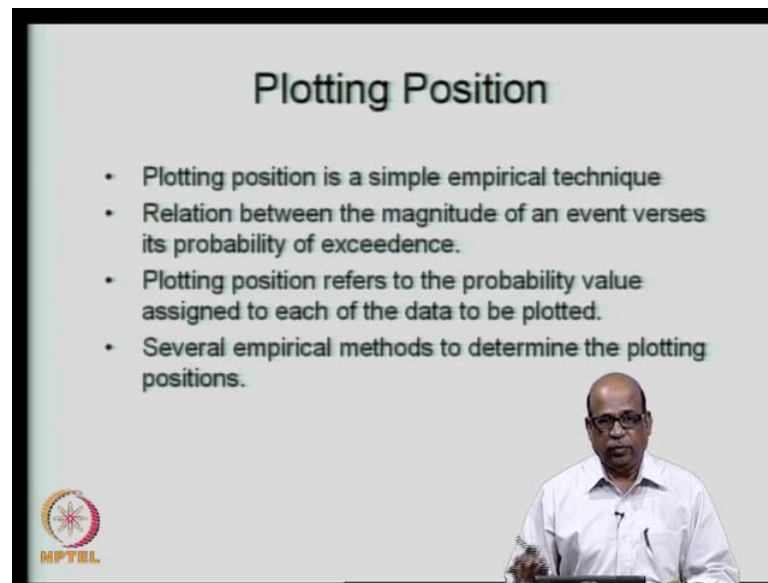
We will now see what we call as the plotting position. Look at the probability papers here. What I said is that, you plot the data on the probability paper. Now, to plot the data on the probability paper, first you must have the data magnitude, which you can plot it on the y axis. But, you also need the corresponding probability values, that is, probability of x being less than equal to x_m or x being greater than equal to x_m and so on.

Now, these are obtained from probability plotting positions. The basis for this is, that you have a small sample of data or even if it is large, comparatively large. It is only a sample. Now, from this sample, you want to estimate the probabilities of a exceedence. That is, probability of x being greater than equal to x given value of x or probability of x being less than a particular value of x and so on.

This we estimate by using the probability plotting positions. We call it as plotting position because, they are not actual probabilities. So, there is a sample that you have, and let us say of the observed data. This observed data, you are now examining whether it is, in fact, drawn from a population belonging to a particular distribution.

Therefore, we use this not as a actual probabilities, but as plotting positions. So that, we can plot it on a probability paper or even on an arithmetic paper. So, typically, we are saying probability of x being greater than equal to x_m , is estimated by such and such a formula. So, we have a number of formulae for plotting position. These are called as plotting position formulae.

(Refer Slide Time: 21:22)



The slide is titled "Plotting Position" and contains the following text:

- Plotting position is a simple empirical technique
- Relation between the magnitude of an event versus its probability of exceedence.
- Plotting position refers to the probability value assigned to each of the data to be plotted.
- Several empirical methods to determine the plotting positions.

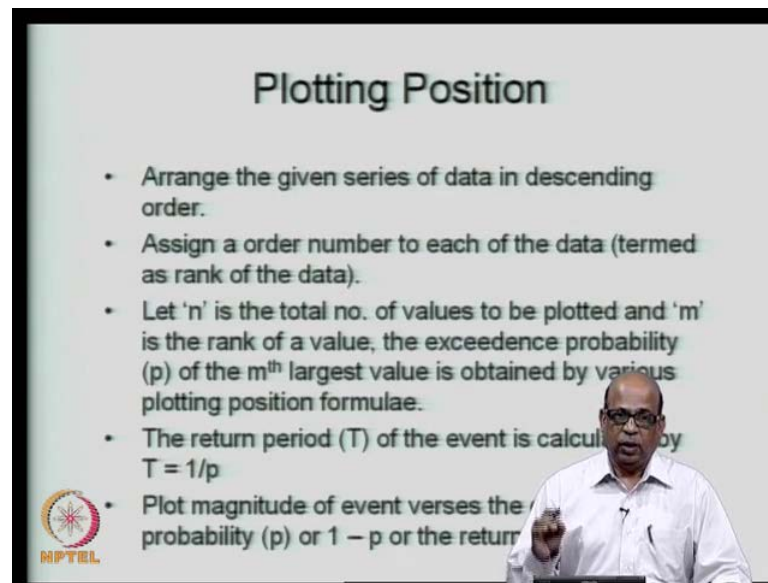
In the bottom right corner of the slide, there is a small video inset showing a man in a white shirt and glasses speaking. In the bottom left corner, there is a logo for NPTEL (National Programme on Technology Enhanced Learning) featuring a stylized sun or starburst design.

So, the plotting position is a simple empirical technique. There is not a rigorous analytical basis for the plotting position. What we do in this is, that we use the fact that there is a relationship between the magnitude of an event versus its probability of exceedence. This exists and this relationship is what we are capturing or what we are depicting through the plotting position.

We say that, the plotting position actually refers to that particular probability value. Although, it is not exactly a probability, it just refers to that particular probability value assigned to each of the data to be plotted. Which means, in the sample, you have several data points. Associated with each of the data point, we provide a probability plotting position. Then, we say that this plotting position, in fact, refers to the probability of exceedence of that particular value.

In hydrology, we use several plotting positions typically. I will go through some of the empirical methods, which provide the plotting positions. Typically, you have, let us say n number of values. Let us say, 50 years of monthly data values, which means 600 values. All your decisions must be based on what you have observed on the field. So, this is observed data.

(Refer Slide Time: 22:58)



Plotting Position

- Arrange the given series of data in descending order.
- Assign an order number to each of the data (termed as rank of the data).
- Let 'n' is the total no. of values to be plotted and 'm' is the rank of a value, the exceedence probability (p) of the mth largest value is obtained by various plotting position formulae.
- The return period (T) of the event is calculated by $T = 1/p$
- Plot magnitude of event versus the probability (p) or $1 - p$ or the return period (T).

MPTEL

So, we are basing all our decisions on the observed data. What we do is that, we arrange the sequence in decreasing order. So, from the highest value first to the lowest value last, we arrange it in a decreasing value. We assign an order or a rank to each of the values, starting with the highest value taking number 1 and the last value, taking rank of m.

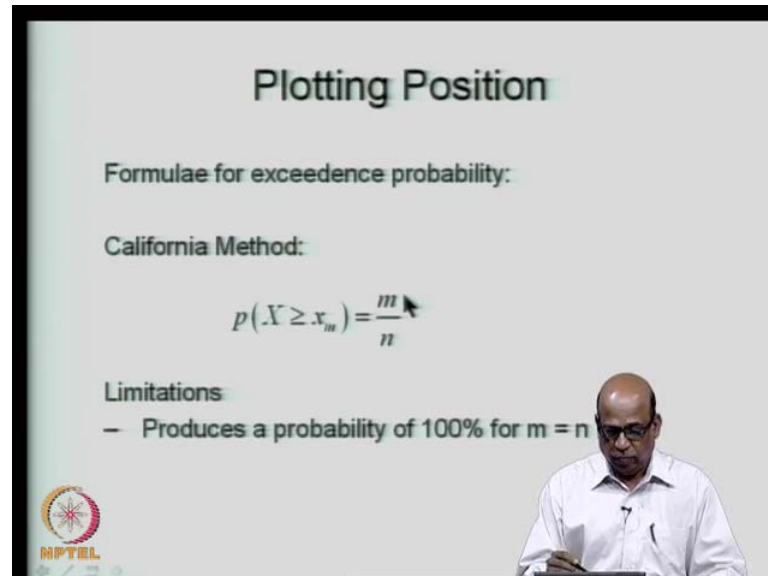
If you have n values, so if you have n number of values and m is the rank of a particular value, then the exceedence probability of the m-th largest value is obtained by various plotting position formulae. That is all. That means, we have assigned to each of the data point, a rank m and n is a total number of values. So, associated with this m, we estimate the probability of exceedence, using both m and n and the return period of the event, of a particular event, is typically calculated by 1 over p, where p is the probability of exceedence.

Once you get the plotting position, we plot the magnitude of event versus the exceedence probability p itself or typically 1 minus p or even the return period. So, the plotting position is used to plot, either p or 1 minus p or the return period, which are all functions of p. So, we typically plot data versus p or data versus 1 minus p, which is probability of non exceedence and data versus the return period.

So typically, these are what we generally use in frequency analysis. Now there are, as I said, several formulae that have been proposed in hydrologic literature for plotting positions. We will just go through them one by one and then, we will use one of the

methods to see how we estimate the plotting positions and then also, see how we plot using the plotting positions on the normal probability paper.

(Refer Slide Time: 25:16)



The slide is titled "Plotting Position". It contains the following text:

Formulae for exceedence probability:

California Method:

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

Limitations:

- Produces a probability of 100% for $m = n$

In the bottom right corner of the slide, there is a small inset image of a man in a white shirt and glasses, who appears to be the presenter. In the bottom left corner of the slide, there is a logo for NPTEL (National Programme on Technology Enhanced Learning).

Similar procedure can be used for any other distribution. So, the California method says, probability, this is a plotting position is simply given by m divided by n , where m is the rank that we have provided to the particular data point. So, x_m is that particular data point and m is the rank associated with that and n is the total number of values.

Now, this has a limitation. It produces a probability of 100 percent for m is equal to n . So, the moment we assign a probability of 100 percent, it means that, we are sure that we have in our data captures, all possible values. Therefore, we are assigning a probability of 100 percent for a particular data point. This becomes a limitation because, we cannot be always 100 percent certain that the data sample that we have is, in fact, representation of the total population.

(Refer Slide Time: 26:24)


Plotting Position

Modification to California Method:

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

Limitations:

- Formula does not produce 100% probability
- If $m = 1$, probability is zero

 MPTEL

Therefore, this becomes a limitation, although, it is a very simple way of assigning probabilities. So, this is how it started the plotting position. Simply, take m by n . A modification was suggested to the California Method, where we do not take m by n , but we take m minus 1 by n .

Now, the moment you put you m by m minus 1 by n , we have taken care of the limitation arising out of assigning 100 percent probability. However, we have introduced another limitation now, that is, m is equal to 1, and the probability is 0. The moment m is equal to 1, it becomes 0. Now, this is not a limitation. The first term formulae, does not produce 100 percent probability. This is not a limitation. However, the second one becomes a limitation. This is a limitation.

That is, when m is equal to 1, you get a probability of 0. Now, probability of 0, again indicates that you have captured all the lower extreme values. Therefore, you are confident that a probability of 0 does exist. That becomes a limitation because, we can never be sure that the sample is, in fact, a complete representation of the population, especially when you are talking about flow probability, flood probabilities and so on, in hydrology.

(Refer Slide Time: 27:46)

Plotting Position

Hazen's formula:

$$p(X \geq x_m) = \frac{m-0.5}{n}$$

Chegodayev's formula:

$$p(X \geq x_m) = \frac{m-0.3}{n+0.4}$$

MPTEL

Then, we have a Hazen's formula, which does a compromise again between, by just taking m by n and taking m minus 1 by n , it takes m minus 0.5 by n . Then, we have a Chegodayev's formula, where it takes m minus 0.3 by n plus 0.4. As I said, these are all empirical methods. They have done several tests on these and then, proposed these particular formulae.

(Refer Slide Time: 28:16)

Plotting Position

Weibull's formula:

- Most commonly used method.
- If 'n' values are distributed uniformly between 0 and 100 percent probability, then there must be $n+1$ intervals, $n-1$ between the data points and 2 at the ends.

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

- Indicates a return period T one year for a period of record for the largest value

MPTEL

Then you have the Weibull's formula, which is the most commonly used formula in hydrology. In fact, the moment you want to use the plotting position, the first thing that

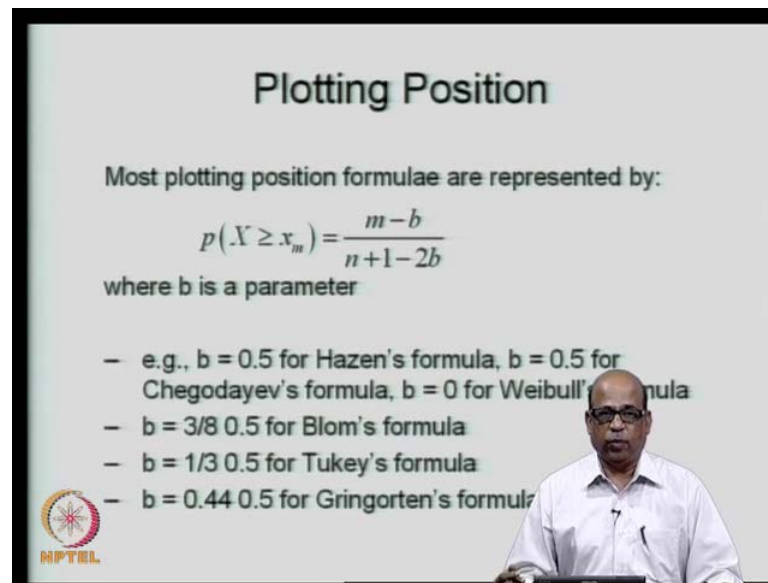
comes to your mind is the Weibull's formula. This is most commonly used for the reasons that I have mentioned presently. So, if you have n values, which are all distributed uniformly, let us say, you have the data sample. Then, if this data is uniformly distributed, then between 0 and 100 percent probability, that is what we are assuming.

That means, you have the data sample and then, between 0 and 100 percent probability, these data are uniformly distributed. That means, there is no probability band, which has more number of data than any other probability band. That is a assumption there. Then, you can get probability of x being greater than equal to x_m is m divided by $n + 1$.

That is, you are leaving out 0 and you are leaving out 1 on the other extreme. Then, remaining $n - 1$ values, $n - 1$ class intervals contain these data in a uniform manner, in a uniformly distributed manner. So, you will have $n + 1$ class intervals. That is, $n - 1$ between the data points and one at the other extreme containing 0 and another at the other extreme containing 1. So, that is how you get $n - 1 + 2$, which is $n + 1$. The probability of x being greater than equal to x_m is then given by m divided by $n + 1$. So, this indicates a return period t , one year longer than the period of record for the largest value. So, you have the largest value in the data and then, you have the largest value should have corresponded to m divided by n . But now, it corresponds to m divided by $n + 1$.

This has several advantages. First of all, it does not produce a probability of 0 because, m starts with 1 and it does not produce a probability of 100 percent because, at m is equal to n , you still have m divided by $n + 1$. Therefore, the Weibull's formula is most commonly used in hydrologic analysis.

(Refer Slide Time: 30:38)




Plotting Position

Most plotting position formulae are represented by:

$$p(X \geq x_m) = \frac{m-b}{n+1-2b}$$

where b is a parameter

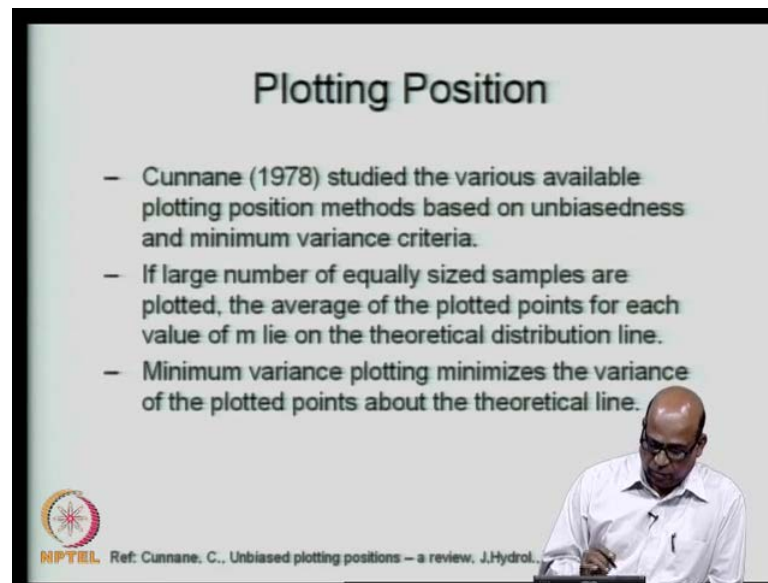
- e.g., b = 0.5 for Hazen's formula, b = 0.5 for Chegodayev's formula, b = 0 for Weibull's formula
- b = 3/8 0.5 for Blom's formula
- b = 1/3 0.5 for Tukey's formula
- b = 0.44 0.5 for Gringorten's formula

 MPTEL

Now, we discussed a few of them. But, there are several other plotting formulas. Most of the plotting formula may be expressed by this simple expression, m minus b divided by n plus 1 minus 2 b, where b is a parameter. So, this gives a general formula for plotting positions. For example, you take Hazen's formula, b is equal to 0.5. So, m minus 0.5 divided by n plus 1 minus 0.5 into 2 2 into 0.5, which is, this goes. So, m minus 0.5 by n, that is what leads to Hazen's formula.

Then, b is equal to 0.5 again, for Chegodayev's formula. For Weibull's formula, you put b is equal to 0, which is m divided by n plus 1 minus 0. So, m divided by n plus 1. Similarly, b is equal to 3.8 for Blom's formula and b is equal to 1 by 3 for Tukey's formula etcetera. This 0.5 does not exist.

(Refer Slide Time: 32:06)



Plotting Position

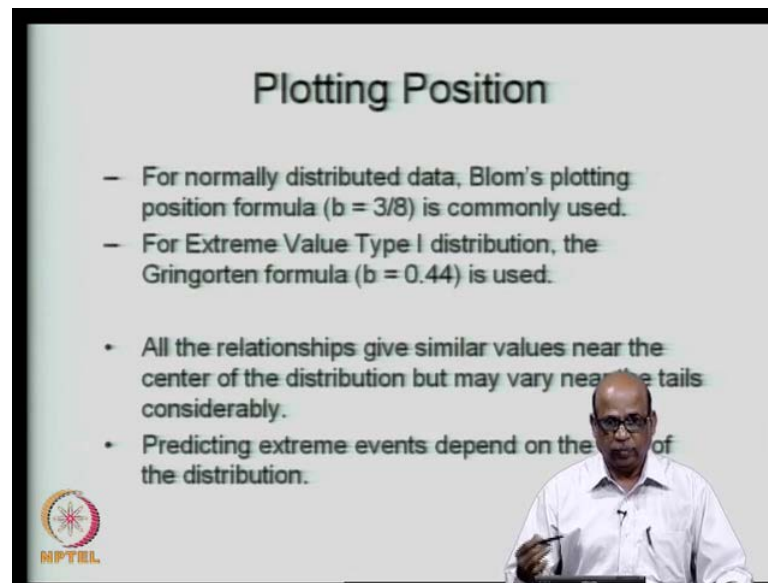
- Cunnane (1978) studied the various available plotting position methods based on unbiasedness and minimum variance criteria.
- If large number of equally sized samples are plotted, the average of the plotted points for each value of m lie on the theoretical distribution line.
- Minimum variance plotting minimizes the variance of the plotted points about the theoretical line.

NPTEL Ref: Cunnane, C., Unbiased plotting positions – a review, J.Hydrol.

Like this, you can generate several formula by using different values of b . Some of them are proposed like this. Now, there is one study which was done long ago in 1978. He uses, that is Cunnane, studied the various available plotting position methods based on unbiasedness. Which of these are unbiased and which of them also lead to minimum variance. He observes, that if large number of equally sized samples are plotted, the average of a plotted points for each value of m lie on the theoretical distribution line.

So, this must be ensured by your plotting position. Then, the minimum variance plotting minimizes the variance of the plotted points about the theoretical line. So, you have the theoretical line and then, you have plotted the data using the plotting position. Then, your plotting position must be such that, among all the plotting positions, it must lead to a minimum variance of plotting.

(Refer Slide Time: 33:26)



Plotting Position

- For normally distributed data, Blom's plotting position formula ($b = 3/8$) is commonly used.
- For Extreme Value Type I distribution, the Gringorten formula ($b = 0.44$) is used.
- All the relationships give similar values near the center of the distribution but may vary near the tails considerably.
- Predicting extreme events depend on the b of the distribution.

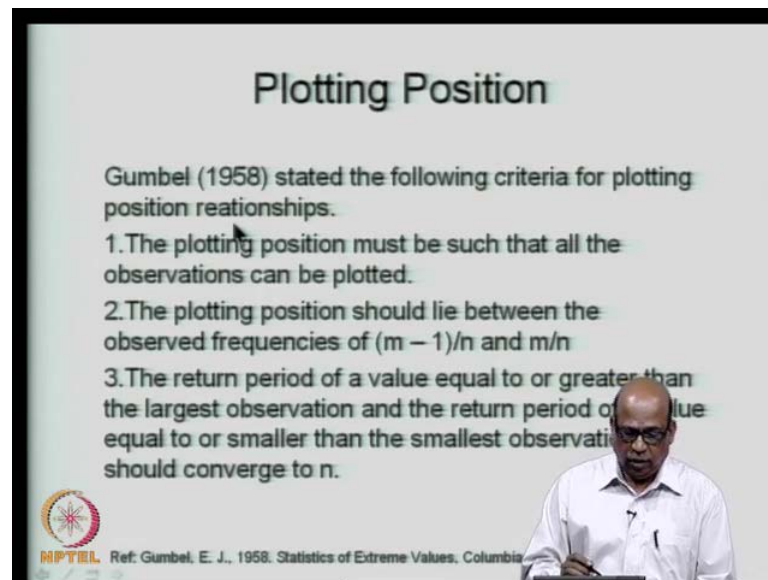
MPTEL

That means, the variance, if you calculate between the observed and the theoretical values, that variance must be minimum. So, this is what we generally use as criteria for choosing a plotting position. But, as a first cut solution to most of the hydrologic problems, you generally use the Weibull's plotting position. For normally distributed data, the Blom's plotting position is commonly used. For extreme value type one distribution, we use the Gringorten formula, which b is equal to 0.44.

Now, whether you use one plotting position or the other, as far as the center of the distribution is concerned, there is not too much of a difference between the plotting positions provided by these various methods. However, when we look at the extreme values, either on the high extreme or on the low extreme, the plotting positions can be quite different from one plotting position to another plotting position.

Therefore, when you are dealing with the extreme values, you must be specifically attentive to which type of these several plotting position formula that is available and which among them you would like to use. Now, this comes from the judgment after looking at the plots of the data on the probability papers. You may see that, at the central point of the plot, most of the data are agreeing with the straight line.


(Refer Slide Time: 35:26)



Plotting Position

Gumbel (1958) stated the following criteria for plotting position relationships.

1. The plotting position must be such that all the observations can be plotted.
2. The plotting position should lie between the observed frequencies of $(m - 1)/n$ and m/n .
3. The return period of a value equal to or greater than the largest observation and the return period of a value equal to or smaller than the smallest observation should converge to n .

 Ref: Gumbel, E. J., 1958. Statistics of Extreme Values, Columbia

But, as you go towards the extremes, the data may be far away from the theoretical straight line. If they are indeed far away from the straight line and if you are using this particular procedure or extreme values, then you must be careful. You should not use such particular distributions. Now, from the Gumbel's study, early in 1950s, we have the following criteria to be met by a good plotting position relationship.

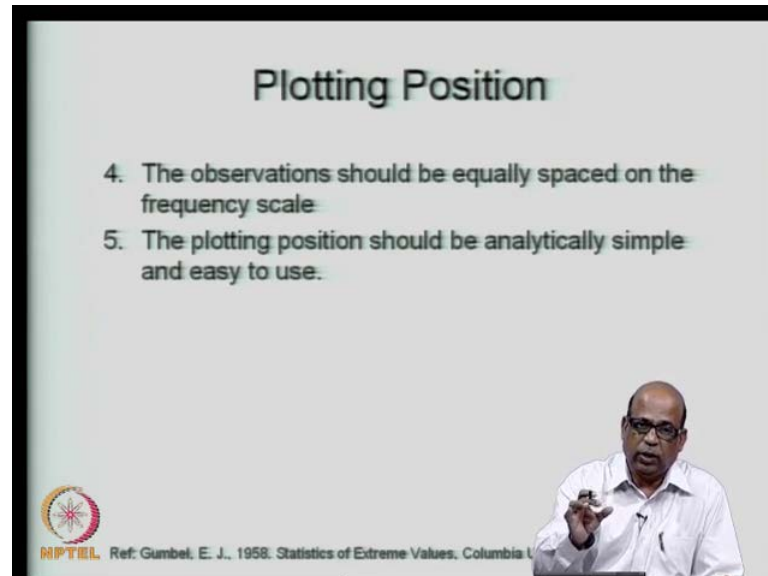
First of all, the plotting position must be such that all observations can be plotted. That is, how you have m , as well as n . You have n number of values and m goes from 1 to n . All of these must contain both m and n . So that, you are able to plot all the observed values. Then, the plotting position should lie between the observed frequencies of m minus 1 by n and m by n . That is, this is the probabilities.

So, as I said, you leave out 0 percent and 100 percent. So, it must lie between m minus 1 by n and m by n . The return period of a value is equal to or greater than the largest observation. The return period of a value equal to or smaller than the smallest observation, should converge to n . Now, this is an important and interesting property of the plotting positions. What do we mean by that? Return period is given by t is equal to 1 by p .

So, you take the largest observation in the data that you have. So, the return period of a value equal to greater than the largest observation. So, get the return period corresponding to that and the return period of a value equal to or smaller than the

smallest value. Both these must converge to n . Now, that is one of the requirements of a good plotting position relationship.

(Refer Slide Time: 37:14)



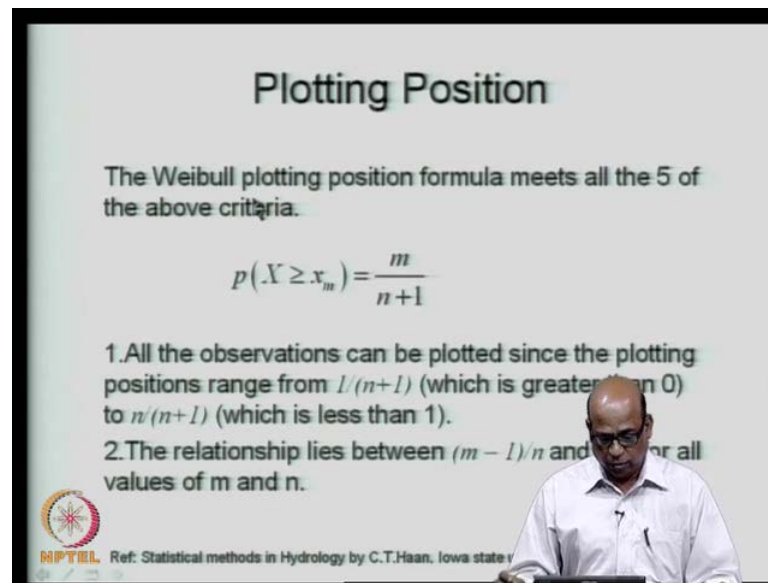
The slide is titled "Plotting Position" and contains the following text:

4. The observations should be equally spaced on the frequency scale.
5. The plotting position should be analytically simple and easy to use.

In the bottom right corner, there is a small inset image of a man in a white shirt and glasses, pointing with his right hand. In the bottom left corner, there is a logo for NPTEL (National Programme on Technology Enhanced Learning) and a reference: "Ref: Gumbel, E. J., 1958. Statistics of Extreme Values. Columbia U."

In the observations, it should be equally spaced on the frequency scale. What do we mean by that? You are saying that, this data should map on to that particular distribution and the distribution of the data themselves in various probability bands must be equal. So, that is what we mean by the observation should be equally spaced on the frequency scale. The plotting position should be analytically simple and easy to use, which is obvious. Because, we want to finally recommend these to the field engineers who are in hydrologic designs.

(Refer Slide Time: 38:10)




Plotting Position

The Weibull plotting position formula meets all the 5 of the above criteria.

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

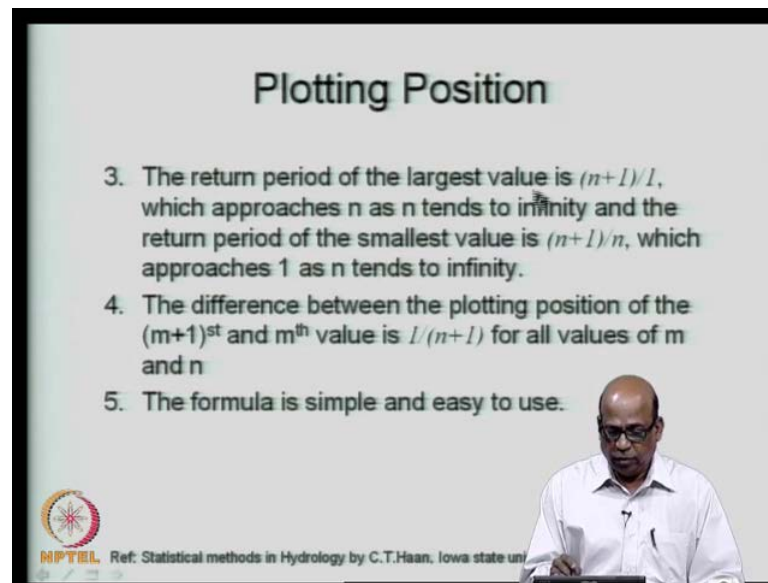
1. All the observations can be plotted since the plotting positions range from $1/(n+1)$ (which is greater than 0) to $n/(n+1)$ (which is less than 1).
2. The relationship lies between $(m-1)/n$ and m/n for all values of m and n .

 Ref: Statistical methods in Hydrology by C.T.Haan, Iowa state v

So, this must be simple. So, these are some of the criteria that Gumbel has brought out for a good plotting position formula. Let us look at the Weibull's plotting position formula. Weibull's plotting position formula is given by probability of x being greater than equal to x_m is equal to m divided by n plus 1.

M is the rank, when we arrange it in the descending order and n is a total number of values. Now, the first condition is that, all the observations we must be able to plot. So, all the observations can be plotted, since the plotting positions range from 1 divided by n plus 1 for the highest value to n divided by n plus 1. So, this is greater than 0 and this is less than 1 and therefore, you can plot all the values.

(Refer Slide Time: 39:10)



Plotting Position

3. The return period of the largest value is $(n+1)/1$, which approaches n as n tends to infinity and the return period of the smallest value is $(n+1)/n$, which approaches 1 as n tends to infinity.
4. The difference between the plotting position of the $(m+1)^{\text{st}}$ and m^{th} value is $1/(n+1)$ for all values of m and n .
5. The formula is simple and easy to use.

MPTEL Ref: Statistical methods in Hydrology by C.T.Haan, Iowa state un

The relationship lies between m minus 1 by n and m by n , because we are talking about m divided by n plus 1, for all values of m and n . Therefore, the second criteria is satisfied. The return period of the largest value is n plus 1 divided by 1. That is, we are talking about m divided by n plus 1. So, it will be, the largest value will have a rank of 1 and the probability associated that will be 1 divided by n plus 1.

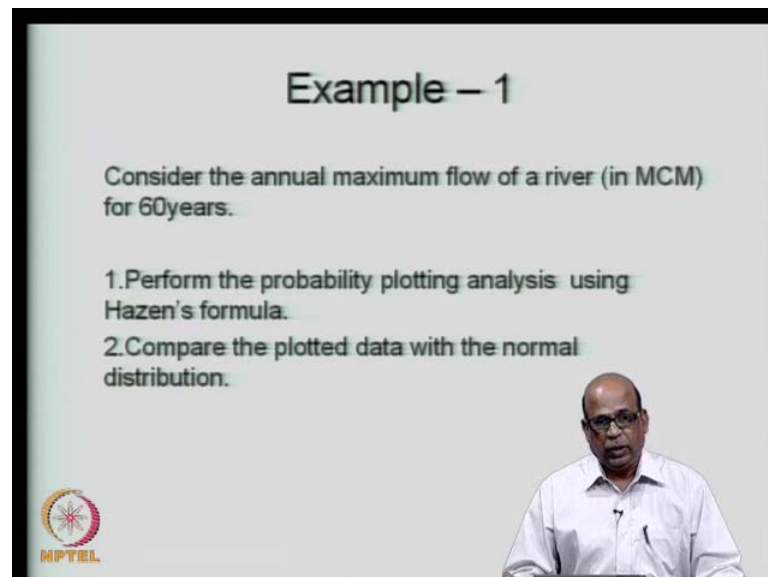
This has to be 1 divided by n plus 1. I am sorry here, 1 divided by n plus 1. For the smallest one, so what happens? I am sorry. This is not. We are talking about the return period. So, return period, as you know is 1 divided by p . Let me explain that. So, return period is 1 by p . So, the return period corresponds to n plus 1 divided by 1. So, if I take out all of this, I will explain this again.

So, we are now talking about the return periods. Not about the probabilities. So, return period is 1 by p . So, 1 by p will be n plus 1 by n , because your p is 1 by n plus 1 and therefore, the return period is n plus 1. So, as n increases, which appears, approaches n as n tends to infinity. So, this will be n plus 1 by 1. It will be almost equal to n as n approaches infinity. Similarly, the return period for the smallest value, that means, we are looking at the last value and the probabilities n divided by n plus 1. Therefore, the return period will be n plus 1 by n , which approaches 1 as n tends to infinity. As n tends to infinity, this becomes 1. That is how it satisfies the condition three of Gumbel's conditions.

Then, the fourth condition is the difference between the plotting position of the m plus first m plus 1 and m -th value is 1 divided by m plus 1 for all values of m and n . What is the fourth condition of Gumbel's? This is the observation. It should be equally spaced on the frequency scale. Now, that is what we are saying here.

The difference between the plotting position of the m plus first m plus 1-th and the m -th value is 1 divided by m plus 1, which remains the same for all values of m and n . That means, the frequency remains the same across different plotting positions and of course, the formula is simple and easy to use.

(Refer Slide Time: 42:12)



The slide is titled "Example - 1" and contains the following text:

Consider the annual maximum flow of a river (in MCM) for 60 years.

1. Perform the probability plotting analysis using Hazen's formula.
2. Compare the plotted data with the normal distribution.


In the bottom right corner of the slide, there is a small image of a man in a white shirt and glasses, presumably the presenter. In the bottom left corner, there is a logo for "MPTEL" featuring a stylized globe.

Therefore, the Weibull's formula meets all the requirements as proposed by Gumbel in 1958. Therefore, the Weibull's formula is most commonly used in hydrologic analysis. We will take a simple example now. We are looking at the maximum flow of the river for 60 years and we will look at the probability plotting analysis. We use the Hazen's formula. I use the Hazen's formula for convenience here because, we use matlab. In matlab, Hazen's formula is what is used for normal distribution. However, you can use Weibull's or any of the other earlier plotting positions formulas that I have mentioned. Then, we will see whether the given data, in fact, can be approximated as belonging to normal distribution.

(Refer Slide Time: 42:50)

Example – 1 (Contd.)

Year	Q (MCM)	Year	Q (MCM)	Year	Q (MCM)	Year	Q (MCM)
1950	1982	1965	1246	1980	2291	1995	1252
1951	1705	1966	2469	1981	3143	1996	983
1952	2277	1967	3256	1982	2619	1997	1339
1953	1331	1968	1860	1983	2268	1998	2721
1954	915	1969	1945	1984	2064	1999	2653
1955	1557	1970	2078	1985	1877	2000	2407
1956	1430	1971	2243	1986	1303	2001	2591
1957	583	1972	3171	1987	1141	2002	2347
1958	1325	1973	2381	1988	1642	2003	2512
1959	2200	1974	2670	1989	2016	2004	205
1960	1736	1975	1894	1990	2265	2005	20
1961	804	1976	1518	1991	2806	2006	73
1962	2180	1977	1218	1992	2532		
1963	1515	1978	966	1993	1996		
1964	1903	1979	1484	1994	1540		




(Refer Slide Time: 43:22)

Example – 1 (Contd.)

- The data is arranged in descending order.
- Rank is assigned to the arranged data.
- The probability is obtained using

$$p(X \geq x_m) = \frac{m - 0.5}{n}$$

- The maximum annual discharge verses the return period is plotted on a normal probability paper.



Now, this is the data. From 1950 to 2009, so, 59 plus 1 and that is 60 years of data. These are the annual maximum discharges in million cubic meters. So, these are the values. This is year and this is the values. What we do is that, we arrange this data in decreasing order. That means, the highest value first and the lowest value last. So, first the data is arranged in descending order.

We assign ranks to each of the data. That means, rank number 1 to the highest value, rank number 2 to the next highest value, etcetera, rank number n to the last value. Then,

we obtain the probability x being greater than equal to x_m . This is the Hazen's formula. m minus 0.5 divided by m , where m is the rank corresponding to that particular value and n is the total number of values.

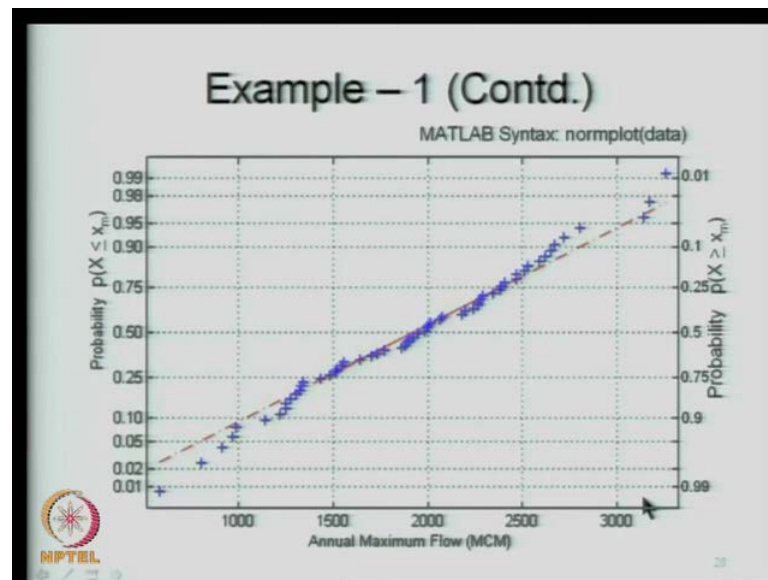
The maximum annual discharge verses the return period is plotted on a normal probability paper. So, this is what we get from the plotting position. Then, we have the given value of the data that is the maximum annual discharge and the return period we can plot. Return period is $1/p$ or we can also plot the probability of exceedence or probability of non exceedence.

(Refer Slide Time: 44:28)

Arranged data	Rank (m)	$p(X \geq x_m)$	Arranged data	Rank (m)	$p(X \geq x_m)$	Arranged data	Rank (m)	$p(X \geq x_m)$
3256	1	0.008	2265	21	0.342	1557	41	0.675
3143	2	0.025	2243	22	0.358	1540	42	0.692
2906	3	0.042	2200	23	0.375	1518	43	0.708
2806	4	0.058	2180	24	0.392	1515	44	0.725
2721	5	0.075	2078	25	0.408	1484	45	0.742
2670	6	0.092	2064	26	0.425	1430	46	0.758
2653	7	0.108	2016	27	0.442	1339	47	0.775
2619	8	0.125	2005	28	0.458	1331	48	0.792
2591	9	0.142	1996	29	0.475	1325	49	0.808
2532	10	0.158	1982	30	0.492	1303	50	0.825
2512	11	0.175	1945	31	0.508	1274	51	0.842
2469	12	0.192	1920	32	0.525	1252	52	0.858
2466	13	0.208	1903	33	0.542	1246	53	0.875
2407	14	0.225	1894	34	0.558	1218	54	0.892
2387	15	0.242	1877	35	0.575	1141	55	0.908
2381	16	0.258	1860	36	0.592	983	56	0.925
2347	17	0.275	1773	37	0.608	966	57	0.942
2291	18	0.292	1736	38	0.625	911	58	0.958
2277	19	0.308	1705	39	0.642	801	59	0.975
2268	20	0.325	1642	40	0.658	511	60	0.992

We can use a probability paper for plotting any of these. So, we take, let us say we pick up the maximum value 3256 and assign rank number 1. Next value, assign rank number 2, etcetera and like this. So, we assign up to ranks 60 in the data. So, it goes from ranks 1 to 60. Once we get the ranks, we can get the plotting position. So, we get plotting position x being greater than equal to x_m using this formula. So, these are the plotting positions.

(Refer Slide Time: 45:10)



Then, we use a normal probability paper, which I had indicated earlier. In this case, we use a normal probability paper. Earlier, I had indicated the probability scales on the x axis. This differs and it is just based on your convenience. You can use either the x axis to indicate your data or x axis to indicate the probabilities. Many commercially available probability papers, we will use x axis as probabilities.

But in this particular case, I am using the matlab program, though the matlab syntax is simply normplot given data. So, data is the vector. You simply give normplot and it uses the Hazen's formula for plotting position and then, generates this plot on the probability paper. So, here we are shown the probability of x being less than equal to x m and here we have shown probability of x being greater than equal to x m. That is what is obtained from this table here.

So, probability of x being greater than equal to x m is what is shown here. So, for example, you have a probability of 0.008. So, you get somewhere here 0.008 and that corresponds to this value. So, on both these axis, vertical axis you can get the associated probabilities. The theoretical line in this case is shown by the red line here. That is, this is the best fit line, straight line for, that is, if this data set is to follow a normal distribution theoretically, then all of these points would have come on this particular straight line.

But, there are some departures. Some of the data are departing from the straight line. As you can see, in the central region of the plot, most of the data are very close to the

straight line. As you go near the extreme, for example, this point, these points here etcetera, as you go near the extremes, the departures are quite significant. So, if you are interested in the central region, then we can reasonably assume that this follows a normal distribution.

Typically, when we want to use normal distribution, we are generally interested in the central region. So, unless you are looking at extremes of, let us say, 1 in thousand year return period, 1 in 500 return period flows and so on, unless you are looking at those extremes, you can reasonably use the normal distribution for this particular data.

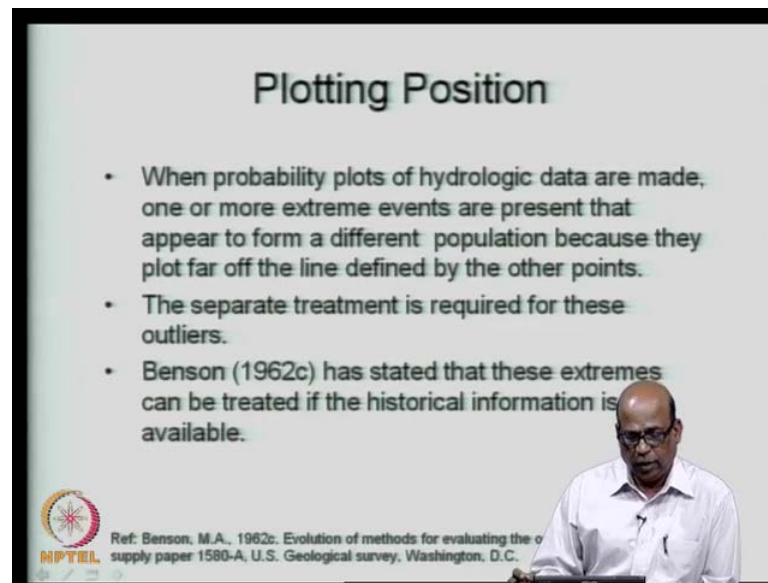
Now, let us say, you look at difference between 10 year return period and 100 year return period. What does 10 year return period correspond to? Let us say, you are looking at t is equal to 10 years. So, t is equal to 10 years and will correspond to the probability 1 by 10, which is 0.1 and t is equal to 100. It may correspond to 1 by 100, which is 0.01.

So, you may be looking at in terms of percentage. You may be looking at 10 percent and 1 percent and the difference between 10 percent and 1 percent. So, this may come to be, let us say your, this is not in percentage. So, you are actually looking at 0.1 and 0.01. So, you are looking at the difference between 0.01 and 0.1.

So, this may appear to be small, when compared to the other scales here, but the values that you get may be significantly different. So, on the graph, you may be tempted to say that there is not too much of a difference. Therefore, I can use this graph also for obtaining 100 year return period, 200 return period and so on because, you do not see too much of a difference in the graph itself here.

However, you must remember that this is a highly non-linear scale and therefore, the values that you get, may be significantly different. Therefore, you must be alert to the situation that, near the extremes, the difference is, if they are significant, the values are significantly departing from the theoretical straight line, then you must be alert to using these for extrapolation purposes, near the extreme values. That is the point that I want to write.

(Refer Slide Time: 50:18)



Plotting Position

- When probability plots of hydrologic data are made, one or more extreme events are present that appear to form a different population because they plot far off the line defined by the other points.
- The separate treatment is required for these outliers.
- Benson (1962c) has stated that these extremes can be treated if the historical information is available.

Ref: Benson, M.A., 1962c. Evolution of methods for evaluating the o
supply paper 1580-A, U.S. Geological survey, Washington, D.C.

MPTEL

So, the plotting position, when the probability plots of hydrologic data are made, typically you know, you may have some extreme values, which may be far away from the straight lines. In this particular case, let us say this value or this value etcetera, they may be far away from the straight lines. In between, we may get a value here. Let us say that, some values may, for example, I may get a value here. Something like this or some value somewhere here.

Now, these are quite far away from the theoretical straight line, when compared to most other values. Most other values, you may say that the departure is only very small random deviation from this straight line. Therefore, it can be accepted whereas, there may be some values which are far away from the theoretical line. These are called as the out layers.

That means, in the data that you have observed, let us say that all of this data were somewhere around this region, but one or two values may be far away from this data. These are called as the out layers. Identification of out layers from a given data is itself an exercise. That means, given the data, do we call this as an out layer or do we include it in the analysis? You may not have just one out layer like that. You may have a series of out layers.

Let us say, some out layer may lie here. Somewhere here and some other value may lie here, etcetera and whereas, most of the other data are all around the line. Now, in that

case, first of all, these are out layers. In such a situation, first identification of the out layers is a problem and then, how do we make use of the out layers in our analysis is another problem. I will not be discussing this in this particular course.

However, you keep in mind that when you get the data like this and then, some of the points are far away, how we treat them is a different topic by itself. You must remember that, you should not simply ignore the out layers. Depending on the type of analysis, the type of use of this analysis that you would like to make, you may have to consider the out layers in different ways.

What I mean by that is, these out layers have actually happened, have actually occurred in the history. Therefore, they are telling you some story. They are giving you some information and this information cannot be simply ignored. So, you must always be alert to the situation that, the out layers are providing some information, which if we ignore, perhaps we may be making too much of an approximation.

So, a separate treatment is required for the out layers. So, we will say here, now there is a good literature available for treatment of out layers. Anyone who is interested in taking these discussions forward, can refer to most of the literature related with out layers. You know, this is classical literature now right from 1960s.

Now, if you have enough historical information; that means, that the data is quite large, data is quite reliable and data has not undergone any changes in terms of homogeneity etcetera, then you can still treat the out layers in a reasonable manner. So, in today's class, essentially what we did is, we started with a probability plotting and continued the discussion on the probability papers. In the last lecture, we had discussed the mathematical way of obtaining the probability papers. Today, we discussed the graphical procedure of obtaining the probability papers, dealing essentially with the normal distribution.

What we do is that, we convert the arithmetic scale into a probability scale, so that, the probability distribution, the c d f plots as a straight line on the probability paper. That is how we obtained the normal probability paper. Similarly, you can also obtain Gumbel's probability paper, log Pearson type 3 probability paper and so on.

Then, to use a probability papers, we need a plotting position. The plotting position, we have introduced several formulae and typically, we use the Weibull's formula for most of the hydrologic analysis, which is given by m divided by n plus 1, where m is the rank of that particular data, when we arrange the data in a descending order.

Why do we use the probability papers? The probability papers, typically are used as the first cut approximation or first cut judgment on whether a given data set that you have, in fact, fits that probability paper or probability distribution or not. If the data, with the associated plotting positions plots as a straight line or nearly as a straight line, then you are reasonably sure that you can use that data sample as belonging to a particular population, which follows that particular distribution, say normal distribution or Gumbel distribution and so on.

You must also be careful in extrapolating the data from the distributions that you so obtain. Let us say that, most of the data in the central region follow the particular straight line very closely. But, at the extremes, you have departures, you have significant departures. Then, you should be vary of doing extrapolations in the extreme regions. So, we will continue this discussion and introduce analytical procedures for testing, whether a particular data set fits a given distribution or not.

That is, we have the data set. What we did through probability papers is that, we constructed the probability papers and then, plotted the data set. Then, saw whether it plots at the straight line and then, made a conclusion or judgment on whether it fits that particular distribution or not. Now, we look at analytical procedures. What do we do in the analytical procedures?

If you want to test whether it fits a particular distribution, say normal distribution, theoretical distribution, we get the expected frequencies from that particular theoretical distribution and we compare it with the observed frequencies arising out of the data. Then, look at how closely, in some statistical sense, how closely the observed frequencies match the expected frequencies. This is what we continue, we discuss in the next lecture. So, thank you for your attention. We will meet again in the next lecture.