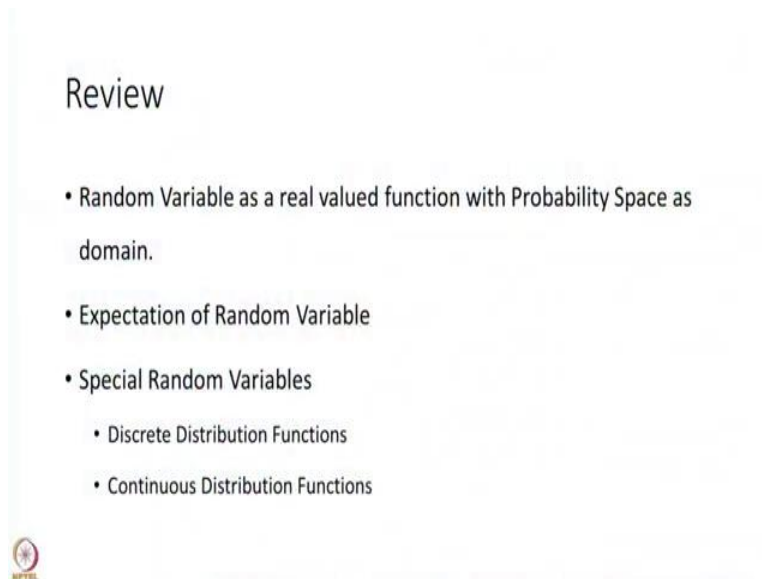


Dealing with Materials Data: Collection, Analysis and Interpretation
Professor. Hina A. Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 36
Probability Plots

Hello and welcome to the course on Dealing with Materials Data. Today we are going to cover the session on what is called Probability Plots. I believe you all have experienced it or have played with it and experimented with it during the R sessions with Prof. Gururajan on Descriptive Statistics, but now the time has come to formally introduce it to you as to what these plots are and what different kinds of plots are available.

(Refer Slid1e Time: 00:57)



So, first let us review what we have done in the past. We introduced Random Variable as a real valued function with a Probability Space as its domain. Then we introduce Expectation of this Random Variable and we also introduced certain special random variables which have a specific models of distribution. So, we introduced two kinds, one is a Discrete distribution functions, in which we covered Bernoulli trials, Binomial distribution, Geometric distribution, Negative Binomial distribution and Hypergeometric distribution.

While in the Continuous distribution function, we introduced Uniform distribution function, then, Normal distribution function, then some derivatives of Normal

distribution such as Chi square, t distribution and F distribution. We also introduced other things distributions such as Lognormal distribution, Weibull distribution, Exponential distribution, I must mention that there are plethora many, many, many distributions available and many, many new distributions are discovered in order to meet today's data requirement.

But we have shown you a few of them, which you come across more frequently and which will be useful to you in your immediate engineering requirements, immediate materials data requirement, of course, as and when we move forward the new distributions that we may require in the further analysis will be introduced in details there.

Now, in all this issue, the question comes that we say that a random sample has been drawn from a distribution or specific distribution? How do we know that it really comes from that distribution? The reality picture is something like this, you have a very large population, which you are trying to study.

For example, if you are studying the yield strength of a particular alloy, which is produced in a factory, in a industry, then how are you going to guarantee that it is going to have this particular yield strength? Well, the method is that you will take a few samples, random samples, you will choose them randomly not a systematically.

And then you will derive certain statistics and you will derive certain values and you would like to see if this distribution and you would have assumed that well theoretically, the yield strength should follow say Lognormal distribution or say Normal distribution, then the question remains is that does sample say that or not?

This is the question we would like to answer in this session and this can be answered through what is known in statistics as Goodness of Fit tests. Goodness of Fit tests are a very theoretical derivation of comparing the data values, data CDF with assumed CDF.

(Refer Slide Time: 04:52)

Outline

- Question: How does one know that given data comes from a specific distribution?
- Goodness - of - Fit tests
- Graphical methods for confirmatory guidelines
 - Probability plots (P - P Plots)
 - Quantile plots (Q - Q Plots)

NPTEL

But there is another method which is called a graphical method, which does not give you a strong proof but it gives you a confirmatory guideline that yes, our assumption that this particular sample comes from this distribution may be correct in this scenario and this is what we would like to cover, we are not going to cover the theoretical Goodness of Fit tests in this course, but, we would like to cover some of the graphical methods for confirmatory guidelines.

We are going to consider primarily two such graphical methods, one is called Probability plots or P-P plots and the other is called a Quantile plot, which is known as Q-Q plots. Now, how do we go about doing this graphical comparison?

(Refer Slide Time: 05:44)

Graphical Comparison

- Assumed that the data is a sample from a particular distribution.
- **Case 1:** Hence, the cumulative distribution function (CDF) obtained from the data should match with the assumed distribution.
- **Case 2:** Similarly quantiles calculated from the data should match with the quantiles of the assumed distribution
- Such matching can be worked out by plotting
 - Data CDF vs. assumed Distribution CDF in **Case 1**
 - Data quantiles vs. assumed Distribution quantiles in **Case 2**



We assume that data sample from a particular distribution, your data what you have got is a sample values coming from a particular assumed distribution. Hence, naturally it means that whatever cumulative distribution function that you will obtain from data should match with the assumed distribution, cumulative distribution function.

Similarly, if they are coming, the sample data is coming truly coming from the one particular assumed distribution, then the quantiles that we have calculated from the data should match with the theoretical quantiles that you would get from the assume distribution.

Such matching can be worked out in two ways, if you are plotting the data CDF versus distribution CDF and if both are equal, they should fall on a straight line. So, this is case one. The other one is if delta quantiles, if you plot against the assumed distribution, theoretical quantiles and they should also match and they should fall on the straight line x is equal to y , this is case two.

(Refer Slide Time: 07:14)

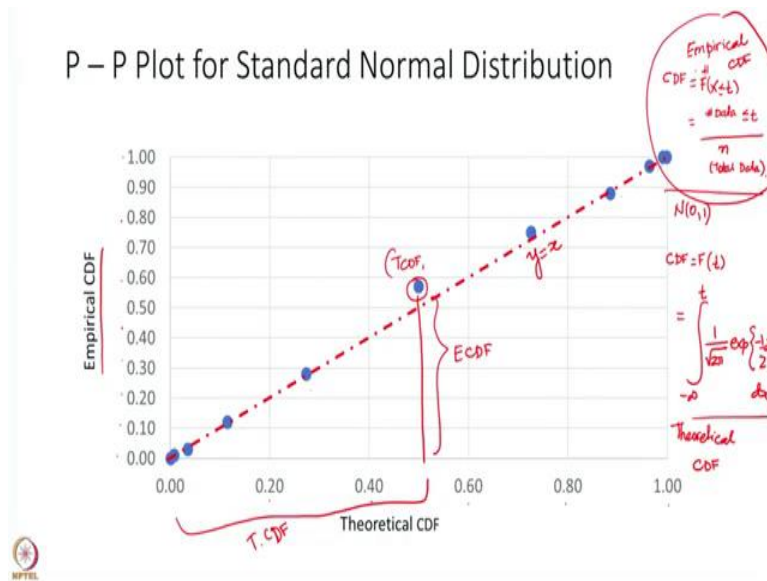
Graphical Comparison

- Assumed that the data is a sample from a particular distribution.
- **Case 1:** Hence, the cumulative distribution function (CDF) obtained from the data should match with the assumed distribution.
- **Case 2:** Similarly quantiles calculated from the data should match with the quantiles of the assumed distribution
- Such matching can be worked out by plotting
 - Data CDF vs. assumed Distribution CDF in **Case 1**
 - Data quantiles vs. assumed Distribution quantiles in **Case 2**



So, here I have detailed described a P-P plot, CDF is calculated from the data and it is called an empirical CDF. Then another CDF is calculated from the assumed distribution and it is called a theoretical CDF. P-P plot refers to plotting the empirical CDF on y axis and theoretical CDF on the x axis. If our assumption is true that the data is truly coming from the, assumed distribution then the points on this plot should fall approximately on y is equal to x line. Otherwise, we should be able to clearly see a mismatch.

(Refer Slide Time: 08:00)



So, let us see the two plots. Now, in this plot what I have done is, I have taken, I have simulated standard normal variates using random number generator, I have simulated about 100 of them and then I have calculated from that data the cumulative distribution function, let us do some recalling here, cumulative distribution function of data is nothing but number of data or let us call it F of X less than or equal to t is nothing but number of data points less than or equal to t divided by n which is total data points.

So, this is called the empirical CDF. So, you calculate the empirical CDF and then from a standard Normal distribution you have a CDF which you can call F of t again then it is nothing but integral minus infinity to t 1 over square root 2 pi exponential minus 1 half x square dx.

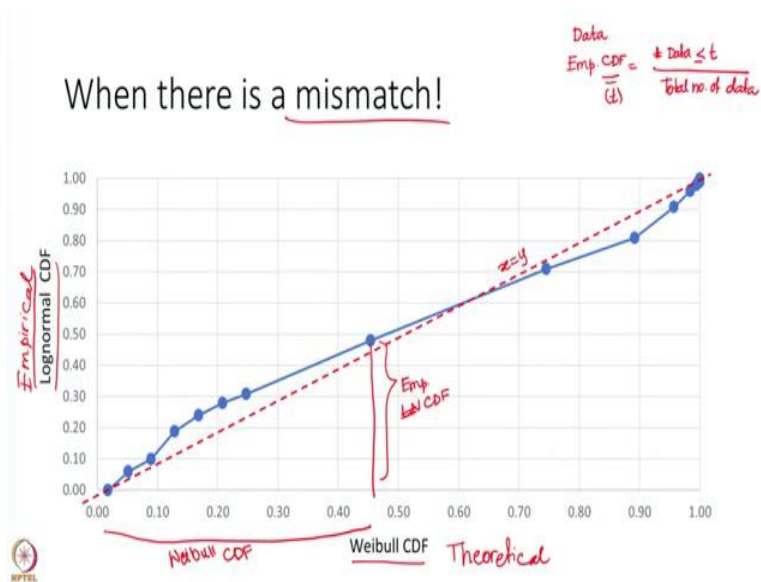
So, this is called theoretical CDF. So, the values of these theoretical CDF on t is given here. This is the x axis, which gives the theoretical CDF, this empirical CDF is plotted on the y axis and that is plotted in, that is shown in here. So, if you take any typical value here.

This says that theoretical CDF which is in here, this is your theoretical CDF value and this is your empirical CDF value. So, this is how all these points are plotted and now, what we say is that if your assumed distribution is the correct distribution for the data that

is your data truly comes from the standard Normal distribution, which in this case, we know because we have generated it randomly.

You can see that it should fall on the line, this is y is equal to x line. So, it should fall approximately on y is equal to x line and you can see that these points are falling and therefore, it says that it conforms graphically that the data points seem to be coming from the standard Normal distribution. Let us take the case of mismatch because sometimes we understand only when we see the matching but more we understand if we see them mismatch.

(Refer Slide Time: 11:58)



Now, here we go here I have generated a Lognormal data by random number generator. So, I have a Lognormal data generated, Lognormal data generated from the random number generator and I am assuming that the data is actually coming from a Weibull distribution.

I have generated the data, so it means that I have a data from a population which actually has a distribution Lognormal, while I have thought that I have actually drawn a sample from a Weibull distribution, I have very purposefully taken these two distributions together is because in the field of metallurgical parameters or properties such as strength property you take a yield strength, you can UTS.

We can consider even the fracture toughness, there is always a question which distribution is closer? And the two competitive distributions for all the strength of the strength property of the material is Lognormal and Weibull. So, here I have taken the Lognormal and Weibull distribution as two competitive distributions.

So, I once again repeat, what we have taken is we have taken a data actually from the Lognormal distribution. But I have assumed or I have believed that the data is coming from the Weibull distribution, then this becomes my theoretical distribution and this becomes my empirical distribution and I have plotted once again.

This shows the Weibull CDF and this shows the empirical Lognormal CDF that is, it is only an empirical CDF it is not really Lognormal I have originally drawn the data from Lognormal. So, please let us understand it clearly, this is a data this is coming the empirical means that it is coming from data. But for my understanding in this course, I have generated this data from Lognormal CDF.

Therefore, I am writing Lognormal otherwise, it is a data. So, just as we did it in the previous case, for a data the empirical CDF is a ratio of number of that points less than or equal to t divided by total number of data. So, this is your empirical CDF for value t and then this is a Weibull CDF which I call a , this is calculated from the Weibull CDF of distribution function and I call it a Weibull CDF.

And now, you see that if I draw this line which is x is equal to y , this is x is equal to y or y is equal to x line, you see that the data is systematically falling above and falling below, there is no random behavior mean that there is it is not an error difference as it happened in the previous case. The data is systematically going above and then systematically going down and therefore, we understand that there is a mismatch with your empirical CDF and you assume CDF, yours assume CDF is far away from what your data says. And remember, data is what we believe, data is what we believe.

(Refer Slide Time: 16:30)

Case 2: Q – Q Plot

$P_1 \ni P[X \leq P_1] = \frac{1}{6}$
 $P_3 \ni P[X \leq P_3] = \frac{3}{6} = \frac{1}{2}$

- Quantiles of Empirical distribution are plotted against the theoretical distribution quantiles.
- If the two distributions are matching then the plot should be approximately along the line $x = y$.
- Otherwise there is a mismatch.

Deciles

Quantiles
 Quantiles: $Q_1 \ni P[X \leq Q_1] = 0.25$
 median = $Q_2 \ni P[X \leq Q_2] = 0.5$
 $Q_3 \ni P[X \leq Q_3] = 0.75$

Now, let us come understand the Q-Q plot. Here again like CDF we take the quantiles of empirical distribution and we plot it against the quantiles of theoretical distributions. You please recall what is quantile, you have understood the quantile in terms of quartile, we are talking about quantiles we have come across the definition of quartiles, quartiles is such that first quartile Q_1 is such that probability of δX less than or equal to Q_1 is 0.25, quartile 2 is such that probability have X less than or equal to Q_2 is half. And Q_3 is a third quartile where probability of X less than or equal to Q_3 is 0.75.

So, if you want to divide the data into four equal parts, your data is here if you want to divide the data into four equal parts, such that the probability from negative infinity to Q_1 , Q_1 to Q_2 , Q_2 to Q_3 and Q_3 onwards are all equal. These are all equal probability and they are all one fourth. This is what is called quartile.

And if you recall this is also known as median, quantile is a general term. So, if you wish to have this probabilities, your data be divided into say five equal parts. Then you will have, if you want to divide it into five equal parts, then you will have each of these, there are six parts here, there is 1, 2, 3, 4, 5 and 6.

So, each one will have a probability one sixth. So, you will have a five values which I call P_1 , P_2 , P_3 , P_4 and P_5 , then this will be called pentile. So if I take a first pentile P_1 is

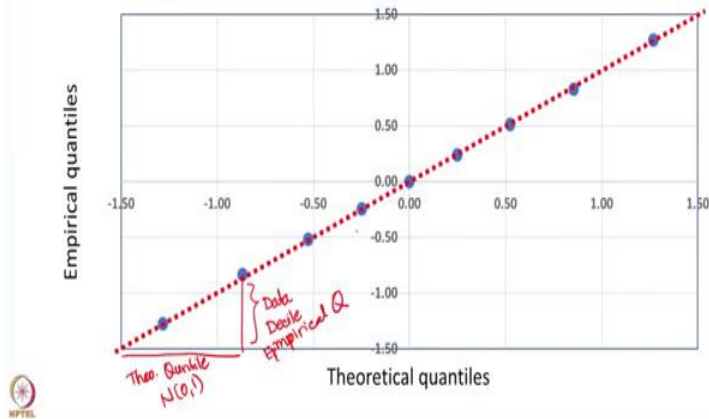
such that probability that x is less than or equal to P_1 is $1/6$ and likewise P_2 is such that probability of x less than or equal to P_2 is $2/6$.

Likewise, you can define so quantiles are a general term. If you want to consider the case of dividing the data into four equal probabilities, you will have quartiles, if you want to, these are not pentiles these hexiles. If you want to divide it into six equal part it will have five points, P_1, P_2, P_3, P_4, P_5 which will divide each data into one sixth. They are called hexiles, you can have deciles, you can have centennials etcetera, etcetera, you have 90 percent data here and there, there are various ways of doing it.

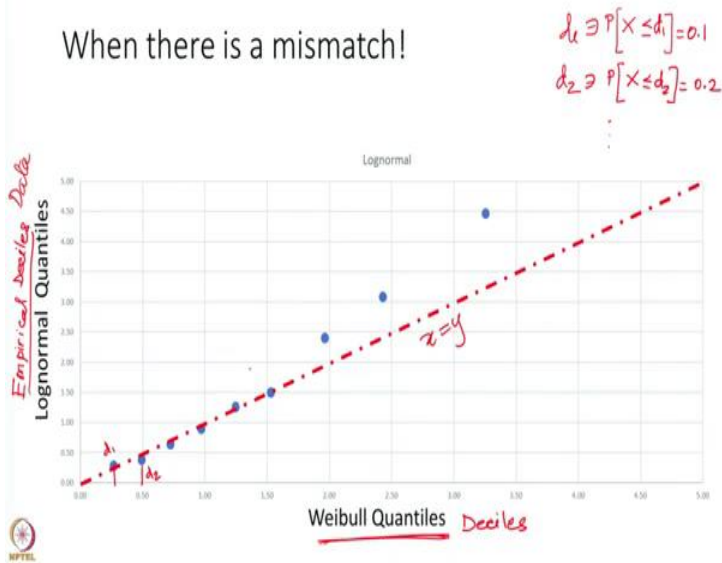
So in this way you can define a quantile. In the example, which I am going to show that there is a matching and there is a mismatch. I am going to consider deciles. I am going to consider deciles it means that the data will be divided into 10 equal probability parts where each part will have a probability you have 0.1, $1/10$ and when I define like this, you remember it is a cumulative probability that I am defining. So I am defining all the values here below this and therefore, it becomes two times $1/6$.

(Refer Slide Time: 21:30)

Q – Q Plot: Standard Normal Distribution



When there is a mismatch!



So, let us see the plots, now. So, if you look at the plot, again I have taken the same standard normal variates generate data generated by using random number generator and I have calculated their deciles. So, there are one, sorry there are 1, 2, 3, 4, 5, 6, 7, 8 and 9 points, remember if it is a decile you are bringing it into 10 parts there will be nine quantiles, nine deciles and these nine deciles I have plotted.

If you take any typical one, if you take any typical one this shows the data decile or as I have call it empirical quantiles and this is theoretical quantiles calculated from normal 0,

1 and I have plotted each value and you can that there is a deviation, I have made the points very big, but otherwise you can look at the center of these points and you know that they are little above on the line or little below etcetera, because I am actually taking random number generated, standard normal, random generator using random generator to generate the standard Normal distribution.

Now, if you look at the mismatch, again I have done the same thing, I have taken the Lognormal distribution to generate random variables of lognormal. In other words, I have used a random number generator and generated lognormal random variables and I calculated quantiles and I am thinking that they have all from Weibull distribution, and I am comparing them so I am doing a little bit of an artificial thing.

But this is to drive the point home, I am following this random number generated values here. So, this is once again empirical deciles and these are Weibull deciles, these are also deciles there are exactly 10. If you look at it, there are 9 data points, 3 and 3, 6, and these 3, 9 and you can see that very systematically it diverts away from the line which is x is equal to y .

And therefore, it says that what you have assumed your data coming from is not the case, your data is coming from some other distribution then the assumed distribution of Weibull. Once again I repeat, that here we are comparing the empirical deciles with the Weibull deciles, this is calculated from data.


So, if I call d_1 as the decile then d_1 is such that probability of data x is less than or equal to d_1 is 0.1, d_2 that is this is the one point, this is d_2 , then this d_2 is such that probability that x is less than or equal to d_2 is 0.2 and likewise. So, this is d_1 , this is d_2 like this I have calculated, the interesting part here is that in the previous case we were matching probability with probability, here we are matching the data value with the data value. So, this is something strikingly different in this case.

(Refer Slide Time: 26:24)

Other Plots

- At times the y axis is scaled to probability scale for different distributions, such as
 - Normal probability scale (paper)
 - Weibull probability scale (Paper)
- The x axis is in the usual numeric scale.
- Probability values are plotted against the value random variable takes
- Here also, if the distributional assumption is correct then the points would fall approximately on the $x = y$ line.

X → Data values
Y = Theoretical probability Scale
x=y
Data



There are other plots, same plots are made in a different way. For example, instead of considering your x axis, instead of considering your x axis as the actual quantiles or actual CDF and y axis also a CDF it says that you take x axis as the data values. So, x axis becomes the data values and y axis is taken as a theoretical probability scale. When there were the computers were not so common, when we studied statistics, there used to be normal probability scale paper available in the market and Weibull probability scale paper available in market.

Now, you do not need as you have done it in the R exercise, you can very easily give a command as to what should be your y axis, what theoretical probability scale you want, and then it plots the data values against the theoretical probability value and again, the matching has to be at x is equal to y.

So, x is in the usual numeric scale showing the data, probability values are plotted against the value of random variables that it takes and here also it falls on a y is equal to x line. So, the exercise that you might have done in the descriptive statistics, R sessions are largely using the probabilities probability scale as y axis and x as a data axis.

(Refer Slide Time: 28:30)

Summary

- Graphical methods introduced to check the distributional assumptions made on the data.
- When Empirical CDF is plotted against the theoretical CDF, the plot is called P – P Plot
- When data quantiles are plotted against the theoretical quantiles, the plot is called Q – Q Plot.
- The same comparison can be made by plotting the empirical CDF on the theoretical distribution probability paper
- In all the above cases, systematic divergence from the $x = y$ line indicate mismatch.



So, let us summarize what we discussed today. We talked about graphical methods to check if the distributional assumptions made on the data are matching or not matching, if you plot an empirical CDF against the theoretical CDF It is called a P-P plot, if you plot a data quantile against the theoretical quantiles then you called a Q-Q plot.

The same comparison can be made by plotting empirical CDF, you can plot the empirical CDF using a probability plot papers and this probability papers, I do not know if the market anymore sells it, but at least you can have it easily on any software package that does the statistical analysis in particular, R has this facility, in all the about cases, it is a matching if it matches x is equal to y line. If there is a mismatch, then that is called indicate a mismatch with your assumption and where the data comes from. Thank you.