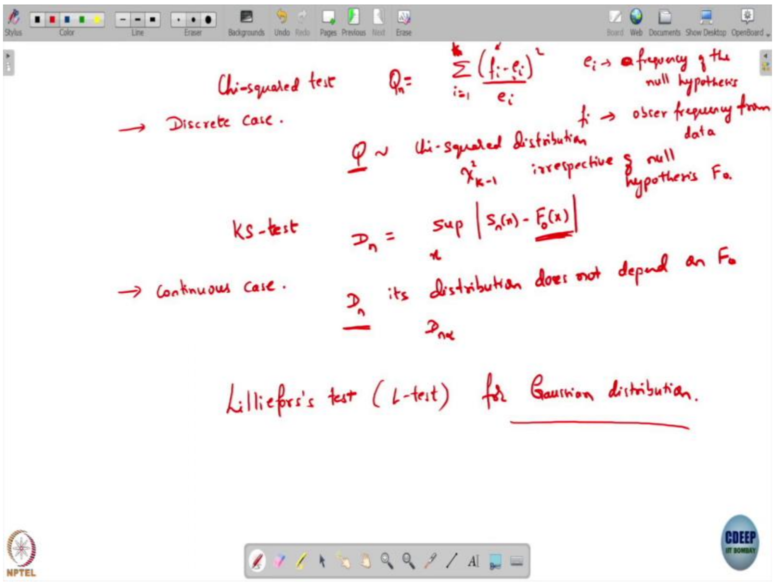
Engineering Statistics
Professor Manjesh Hanawal
Department of Industrial Engineering and Operational Research
Indian Institute of Technology, Bombay
Lecture 58
Lilliefors's Test and Explorator Data Analysis,
Q-Q Plot and P-P Plot

So, so far, we looked into the Kolmogorov-Smirnov test, which we argued that it is a distribution free test, because the test distribution of test statistics does not depend on the underlying distribution of the null hypothesis.

(Refer Slide Time: 0:30)



Chi-squared test

$Q_n = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$

$e_i \rightarrow$ frequency of the null hypothesis

$f_i \rightarrow$ obser frequency from data

→ Discrete case.

$\underline{Q} \sim$ Chi-squared distribution χ^2_{k-1} irrespective of null hypothesis F_0.

KS-test

$D_n = \sup_x |S_n(x) - \underline{F_0(x)}|$

→ Continuous case.

$\underline{D_n}$ its distribution does not depend on F_0

D_{nx}

Lilliefors's test (L-test) for $\underline{\text{Gaussian distribution.}}$

So, let us let us recap this what we said let us look into the 2 things, for first thing which looked into Chi-square test where we said that we are going to look into the statistics which is given by equals to 1 to n, f i minus e i square divided by e i, where e i is the expected, sorry the frequency of the null hypothesis, and f i's were the observed frequency from data.

So, in computing this all we needed is f i's, and this required us to specify the complete distributions so that we can compute e i and in case we do not know that we estimated this, by first estimating the parameter, and then of the distributions, and from the distributions, we computed the frequencies of the classes.

And then the good thing we are able to show that this Q approximately follow Chi-square distribution. In particular, when we said that when the distribution is completely specified, we said that and we have to, we have k classes this Q was exactly Chi-square distribution with k minus, 2 degrees of freedom.

Now, in KS test, and again there were 2 things here we said that this is for discrete case. In the KS test, our maybe we could have put substituted n, put subscripted n also here, because this is based on, or not n samples, or maybe k just to indicate that there are k classes and so, here it was k not exactly n, because we were only making k comparisons, one corresponding to each class. In the KS test we have this, and here k, Q was Chi-squared distribution, distributed and independent of the underlying. So, maybe independent, irrespective of our null hypothesis distribution F not X.

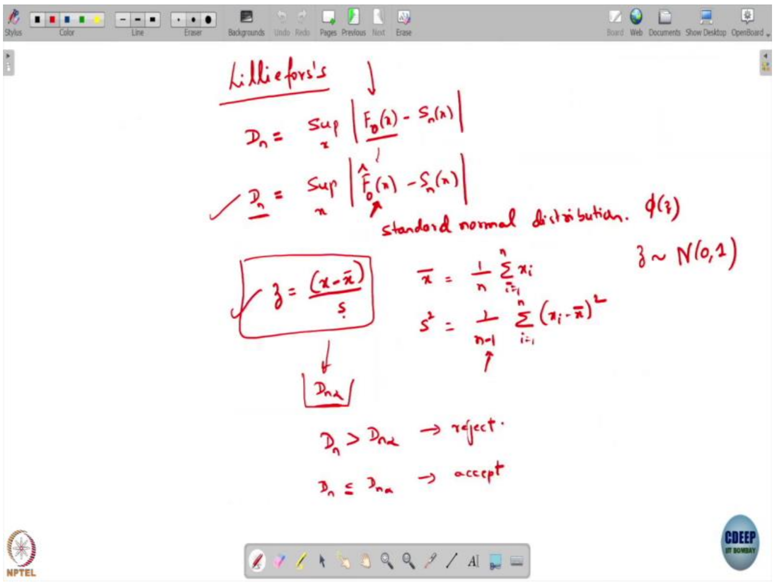
And here, we also said that this D n, in the case of Chi-square test, our statistics we could explicitly establish that it is Chi-square distribution in D n, all we could argue is that, it is its distribution does not depend on. So, we did not explicitly obtained the distribution of D n, all we said is, one can compute them and they are available in the form where D n alpha values will be specified from which we can obtain the Alpha critical points of these distributions.

And again, here it was for continuous distribution. So, notice that the KS test required us to specify completely the distribution that we are going to test. And in case we do not know all the parameters of this distribution, but only the shape is specified one has to estimate those parameters and then use them in this distribution F not x.

However, once you estimate that and plugin your, and use it for your hypothesis distribution, then the computation of the distribution of D n becomes complicated because of that, it is not easy to establish the tables for it. And in fact, it so happens that if you pretend that whenever there is no parameters specified of your distribution, but if you even if you plug in the estimated one and pretend that that is the actual, that is the true value and continue to use your D n tables, the values you are going to obtain can be very conservative. And because of that, you may end up making more errors in accepting or rejecting your distributions.

To somewhat overcome this issue, we have another test called as let me make sure that I can spell it correctly, we will be looking into Lilliefors's test, for short I am going to just call it as L-test. To check my hypothesis whenever my underlying hypothesis that I need to test is Gaussian. Here all I am specified is the, I need my null hypothesis is just to Gaussian distribution, and we have not been told what is the mean and variance. So, then I need to check whether my data samples I how they follow Gaussian shape, how to go about this.

(Refer Slide Time: 07:57)



Lilliefors's

$$D_n = \sup_x \left| F_0(x) - S_n(x) \right|$$

$$\checkmark \; D_n = \sup_x \left| \hat{F}_0(x) - S_n(x) \right|$$

standard normal distribution. $\phi(z)$

$z \sim N(0,1)$

$$\boxed{z = \frac{(x-\bar{x})}{s}}$$

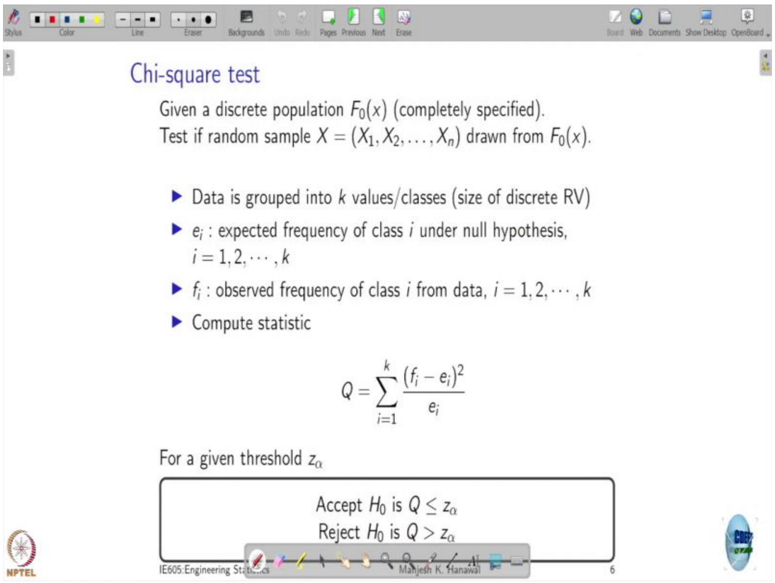
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$D_{n\alpha}$

$D_n > D_{n\alpha} \rightarrow$ reject.

$D_n \leq D_{n\alpha} \rightarrow$ accept



Chi-square test

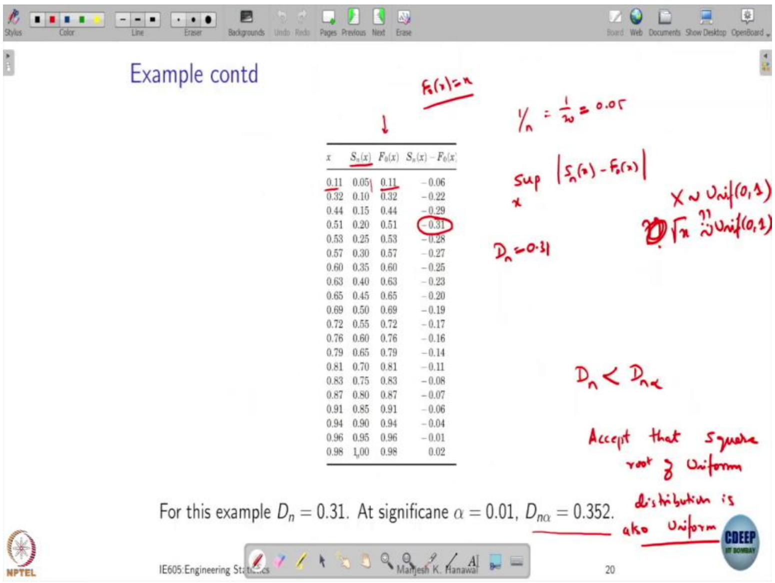
Given a discrete population $F_0(x)$ (completely specified).
Test if random sample $X = (X_1, X_2, \ldots, X_n)$ drawn from $F_0(x)$.

- Data is grouped into k values/classes (size of discrete RV)
- e_i : expected frequency of class i under null hypothesis, $i = 1, 2, \cdots, k$
- f_i : observed frequency of class i from data, $i = 1, 2, \cdots, k$
- Compute statistic

$$Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

For a given threshold z_α

Accept H_0 is $Q \leq z_\alpha$
Reject H_0 is $Q > z_\alpha$



Example contd

$F_0(x) = x$

x	$S_n(x)$	$F_0(x)$	$S_n(x) - F_0(x)$
0.11	0.05	0.11	−0.06
0.32	0.10	0.32	−0.22
0.44	0.15	0.44	−0.29
0.51	0.20	0.51	−0.31
0.53	0.25	0.53	−0.28
0.57	0.30	0.57	−0.27
0.60	0.35	0.60	−0.25
0.63	0.40	0.63	−0.23
0.65	0.45	0.65	−0.20
0.69	0.50	0.69	−0.19
0.72	0.55	0.72	−0.17
0.76	0.60	0.76	−0.16
0.79	0.65	0.79	−0.14
0.81	0.70	0.81	−0.11
0.83	0.75	0.83	−0.08
0.87	0.80	0.87	−0.07
0.91	0.85	0.91	−0.06
0.94	0.90	0.94	−0.04
0.96	0.95	0.96	−0.01
0.98	1.00	0.98	0.02

$1/n = \frac{1}{20} = 0.05$

$$\sup_x \left| S_n(x) - F_0(x) \right|$$

$X \sim \text{Unif}(0,1)$

$\sqrt{x} \overset{??}{\sim} \text{Unif}(0,1)$

$D_n = 0.31$

$D_n < D_{n\alpha}$

Accept that square root of Uniform distribution is also Uniform

For this example $D_n = 0.31$. At significane $\alpha = 0.01$, $D_{n\alpha} = 0.352$.

So, now, that is where Lilliefors's contributions come into picture. So, now the suggested modification here is here we have D n in the KS statistics we have. So, here it worked on all points X, one possibilities, we can instead of this, we can use an another form of this, which is I am going to denote as F n and this is now taken to be standard normal distribution.

That we have earlier denoted as phi of z. Now, before I use it, we are going to do some transformation on my data. Instead of looking into all the X's that I have, we will be looking into the transformed one. And how I am going to transform is like this. And here X bar is simply the mean, empirical mean of the samples I have observed, S square is my standard empirical variance.

And this is like i equals to 1 to n, X i minus X bar whole square. So, notice that I am, as we did earlier, I am using the denominator n minus 1. So, to ensure that my variance estimator is normalised. And here S is, I am going to take it square root to get my empirical standard deviation. So, now we know that roughly z, we have basically centralised and normalised. So, this is roughly going to behave like a normal distribution with 0, 1.

And now, I am going to compare this with, basically the standard normal distribution. So, now, everything remains the same. And again, the distribution of this can be computed not in explicitly form, but maybe through some empirical evaluations. But the good thing is this, even in this case, after doing all this transformation, and comparing with the standard normal distribution, it is still distribution free, because I do not need to worry about what is the underlying distribution in computing or evaluating the distribution of D n.

So, I think Lilliefors did extensive computation of the tables for this D n. And again, the D n alpha tables are available after doing this transformation. And then again, we can do our test using the similar criteria we have earlier. So, when I want to compute, calculate, or like, I want to check whether my distribution of the data follows a Gaussian distribution, we do this transformation and then you compute the statistic.

And for this statistic, we already have that tables, that gives me the critical, the thresholds for various significance values, then whatever D n value I have I am going to compare it, whether this holds or not. If this holds, then reject this to be following Gaussian distribution. And then if it is less than this value, then accept it.

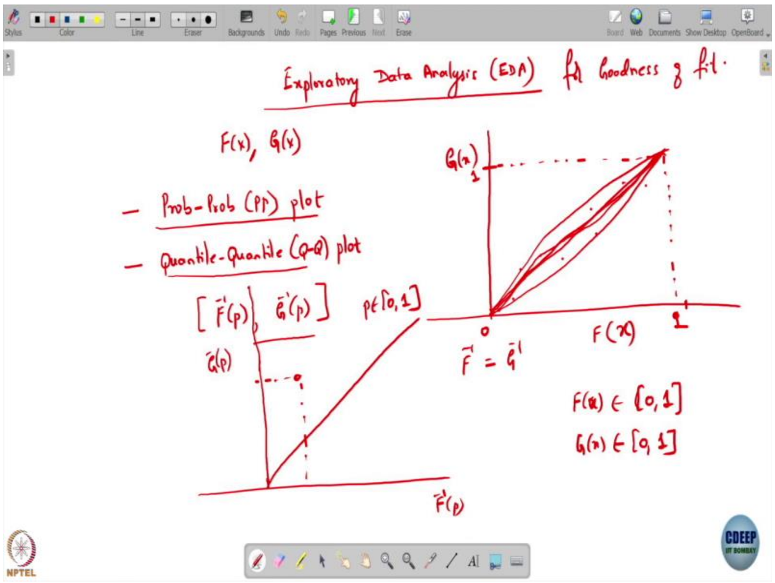
Well, just let me check whether did you say greater than or equal to, equal to was include or it was strictly great. I think we made strictly great for rejection, and accept with less than or equals to, fine. So, this is the summary of Lilliefors's test. And the example is one can follow the same steps as we did here.

The only thing that needs to be done is we have to use this transformation before we applying that, and use the standard normal distribution in the definition of D n and again, use the D n alpha value which are computed for these specific statistics. So, simply when you are using when after you do this transformation, and you when you use this standard normal distribution here, just do not use the tables that you used here to when you try to apply the KS statistics.

So, use a separate tables that are available for this Lilliefors's test, and based on whether it is larger or not you can accept or reject your hypothesis that your distribution follows Gaussian distribution, or not. Now, this is all we computed, we did, we studied several methods about weather my datas follows certain distributions, and if the distributions are kind of fixed, we are only not sure about the parameters.

Then we also studied various hypothesis test for that. So, we studied various parametric and non-parametric method. But often before we get into this parametric and non-parametric method. So, one can do simple visualisation of the data itself and see that whether the data is following my hypothesis distribution. So, for that there are various methods, I am going to quickly discuss two of them.

(Refer Slide Time: 15:16)



Exploratory Data Analysis (EDA) for Goodness of fit.

$F(x), G(x)$

- Prob-Prob (PP) plot
- Quantile-Quantile (Q-Q) plot

$[F^{-1}(p), G^{-1}(p)]$ $p \in [0,1]$

$F^{-1} = G^{-1}$

$F(x) \in [0,1]$

$G(x) \in [0,1]$

So, this is something called exploratory data analysis, so let us say you have been given two distributions, CDF, G and C, and you want to check with our how close, or how similar, or how dissimilar they are? To do this, we are going to look into two possible, or we are going to study two things, one is called probability-probability plots, also often called plot, and another is called quartile-quartile plots, sorry quantile-quantile plots, called Q-Q plots. So, this is what we are trying to do is expose the trade rate analysis for goodness of fit here. So, let us try to look into what is this probability-probability plot.

So, let us say you have this x, for every possible value of x you can take F of x here so, the F of x, the range was going to be we know that F of x range is going to be between 0, 1. And so, is G of x. So, on the x axis you are going to take all possible values of F of x, and on the y axis you are going to look into all possible values, that y axis represent all possible values so, let us say this is 0, 1.

So, this is like you are looking into this box here, because I do not need to go anywhere beyond sorry, this is 1, and do not unnecessarily. So, let us draw one line with slope 45, I would say from this figure you can see that this is not exactly a square shape, it is looking more little rectangular, but assume this is just like square here. So, I have drawn a diagram which has a slope of 45 degrees.

Now, you can plot G of x versus F of x and maybe I do not know like for all possible values of x you may get something like this, and for a given set of. So, if you are a G and G and F are completely specified to you maybe you will get like some continuous line for all possible x, you may get or you may get like something like this, sorry, I made a mistake, this is like CDF, whatever like it can go up down whatever, and then or you may get something like this. So, now clearly if your line is, your plot of F versus G lies very close to your 45-degree line, then you can claim that your data is following up, you can say that these 2 distributions are similar.

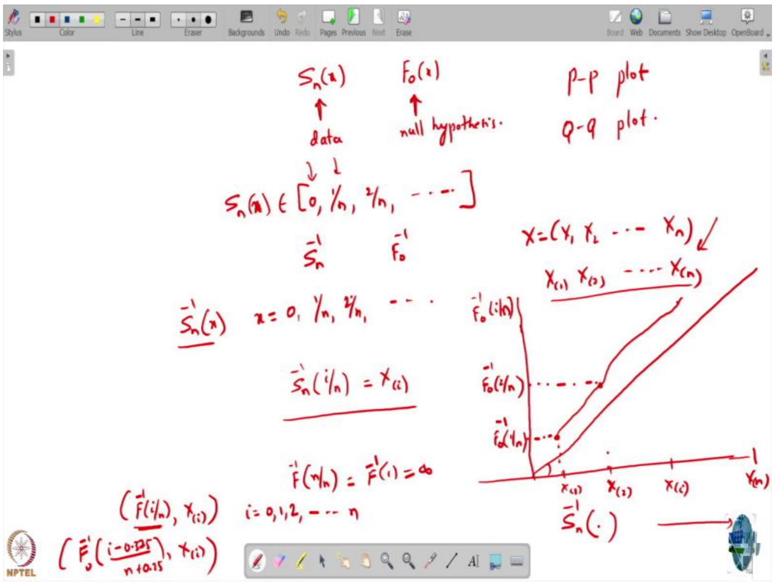
And if it is too much off from your 45 degrees, then it kind of gives an indication that maybe the these 2 distributions are different. Maybe you need to explore this data more, or at least do further systematic analysis to actually say that, okay you can declare that these are very different.

Similarly, instead of going for P-P plot, one can also do Q-Q plot. So, what is Q-Q plot doing? Instead of simply plotting F and G we can plot F inverse of P and the G inverse of P and here P has to be between 0 and 1. I noticed that now the range of F inverse P and G inverse can be like entire real line. Now, again we can do the same thing, P is here. So, for all possible of P values where P is between 0, 1.

You get the values for a particular, let us say this is a particular, for a particular P let us take a particular P and this is one value corresponding to that, then I will get like this is corresponding to let us say this like F inverse P and like this is like a G inverse P, I will get this one value for a particular value, and like that you do it for all possible values of P ranging between 0, 1.

So, for every possible value of P, you will going to get a different lines and you see that if F inverse is same at G inverse you are again going to get a 45 degrees line, which is going to pass through. And if they are not, there then you are not going to get a line which is going to be close or overlapping with these 45 degrees, and based on that you can based on whether you are line is going to be closer, or how far from you are 45 degrees line, you will get a sense of whether your data is going to follow the given distribution. Or like in this case, you will get a sense of whether these 2 distributions are same or not.

(Refer Slide Time: 22:37)



$S_n(x)$ ↑ data

$F_0(x)$ ↑ null hypothesis.

P-P plot

Q-Q plot.

$S_n(x) \in [0, 1/n, 2/n, \ldots]$

S_n^{-1} F_0^{-1}

$S_n^{-1}(x)$ $x = 0, 1/n, 2/n, \ldots$

$S_n^{-1}(i/n) = X_{(i)}$

$F^{-1}(n/n) = F^{-1}(1) = \infty$

$X = (X_1, X_2 \ldots X_n)$

$X_{(1)}\, X_{(2)} \ldots X_{(n)}$

$(F^{-1}(i/n), X_{(i)})$ $i = 0, 1, 2, \ldots n$

$\left(F_0^{-1}\left(\frac{i-0.375}{n+0.25}\right), X_{(i)}\right)$

Using this we can now think of comparing what we have, I have this S n x which is computing from my data, and F 0 x which is from my null hypothesis, can compute. Often in this case instead of the P-P plot, Q-Q plot become easier to plot because of the properties of S n x. So, what do we know S n i first of all takes value, discrete values like 1 by n, 0, 1 by n, 2 by n like this.

So, because of this I need to when I have to look into the Q-Q plot I have to look into S n inverse and F inverse, I only need to calculate this S n inverse at this discrete point, that is I need to calculate S n inverse at like at x, where x is these quantities 0, 1 by n, 2 by n. And moreover, once I have this given data X 1, X 2, X n and I have this order them, and I have this order statistics, I know that S n inverse at let us say some i by n is exactly close to X of i.

We know this already. This is the property of your empirical distributions. So, because of this, let us say you are plotting your Q-Q plot. You are X axis let us say if you are going to plot you are going to make X axis corresponding to empirical distributions. So, these points are going to be simply X of 1, X of 2 like this.

And so, this is like corresponding to 1 by n, and now on the. So, now, whatever this corresponds to now on this, this has this point corresponds to where your F of 0 is let us say 1 by n and here this one corresponds to maybe let us say if this corresponds to you will have this point to be let us say F of 0 by 2. So, like this. So, all you need to do is on your Y axis, you have the points F 1 by n and F 0 2 by n, and the corresponding points on your X axis is going to be X of 1, X of 2.

And now, you need to see you have these points and let us say this is corresponding to F 0 phi by n. Let us say this is just you need to take the point through them, and see how close it is to use 45-degree line. So, this gives you when you are looking into the Q-Q plot against your empirical distribution, when you are comparing Q-Q plot of your empirical distribution with that after null hypothesis, you are kind of X axis you know which points you have to look into, on the y axis you know which points you have to look into.

So, you know what is your curve, all you need to do is check whether that car has a 45 degrees slope, if it has a 45-degree slope then are very close to that, is a good indication that this must this your data is following null hypothesis distribution. Otherwise, you need to do some more confirmatory tests like what you have done before.

One small issue with this is like we know that, then we have this X of n here. So, X of n this will corresponds to the point F of n by n. So, this should be, we should be inverse here. Now, this will correspond to when you have X of n here, this will corresponds to F inverse n of n, and we know that F inverse of 1 is infinity.

So, because of this at as n is when we are exhausting all the endpoints, when you are reaching my last point of my order statistics, my F inverse is tending infinity. So, often to overcome this issue, the packages will do slightly different things. So, what we have here is actually so, what we have is when we are going to plot this what we are basically doing is F inverse of i by n and X of i, and we are plotting them like this is for we are plotting the i equals to 0, 1, 2, to n.

And here when i becomes n, this point is at infinity the Y axis is infinity. So, most of the packages to overcome this issue, they will look at a slightly modified version of this. I am directly taking it from this like this maybe they will do some small alteration to this by taking i 0.375 divided by n plus 0.25, and then do this.

So, when i is becoming n, you need to do some corrections so that you are not hitting infinity on your Y axis. So, one of the some, I am just mentioning that there could one of the possible correction is to do this, so, that you can still observe some value on the Y axis. And this is just one of the examples, you may do various different combinations, so that you can show something still on in your plot.

So, often even though we discussed this exported data analysis at the end, maybe this is the first thing you want to do, you want to just draw your Q-Q plots, P-P plots and see how close they are compared to the 45 degrees lines. And if they are pretty close, you get a good confidence that okay, your data is as per your null hypothesis.

Otherwise, maybe data is not enough. So, you need to do some more test. And that is where he can revisit all the hypothesis testing, we did when you our distributions are parameterised, and or you want to use a full knowledge of the distribution. Or, you can use no nonparametric methods when you do not want to use the knowledge of your distributions to make this test. So, with this, we will conclude this. Thank you very much.