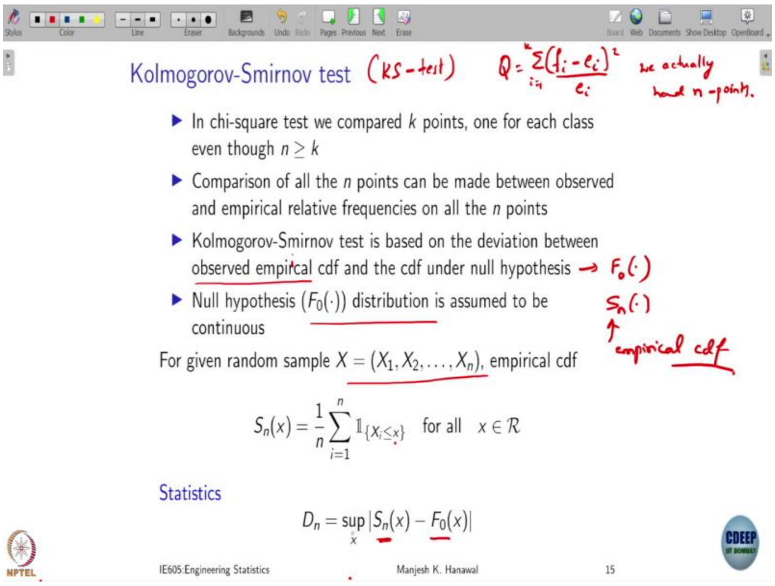
Engineering Statistics
Professor Manjesh Hanawal
Department of Industrial Engineering and Operational Research
Indian Institute of Technology, Bombay
Lecture 57
Kolmogrov-Smirnov Test

After seeing this, how to apply our Chi-square test, let us now start looking into another test, Kolmogrov-Smirnov test.

(Refer Slide Time: 0:24)



Kolmogorov-Smirnov test (KS-test) $Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$ we actually had n-points.

- In chi-square test we compared k points, one for each class even though $n \geq k$
- Comparison of all the n points can be made between observed and empirical relative frequencies on all the n points
- Kolmogorov-Smirnov test is based on the deviation between observed empirical cdf and the cdf under null hypothesis → $F_0(\cdot)$
- Null hypothesis ($F_0(\cdot)$) distribution is assumed to be continuous

$S_n(\cdot)$ ↑ empirical cdf

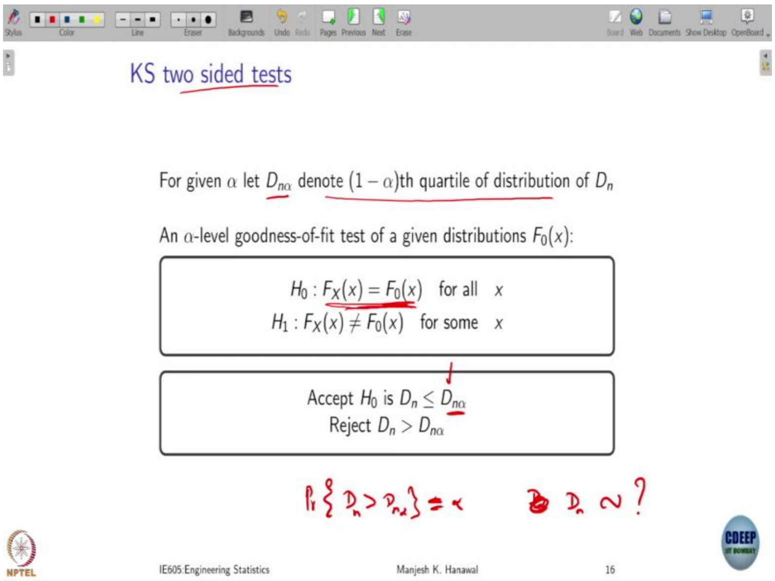
For given random sample $X = (X_1, X_2, \ldots, X_n)$, empirical cdf

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \quad \text{for all} \quad x \in \mathcal{R}$$

Statistics

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

IE605:Engineering Statistics Manjesh K. Hanawal 15



KS two sided tests

For given α let $D_{n\alpha}$ denote $(1-\alpha)$th quartile of distribution of D_n

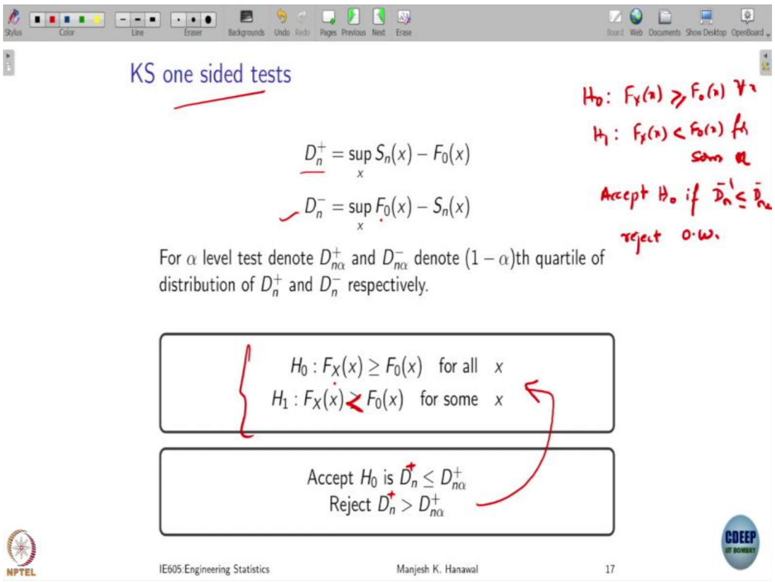
An α-level goodness-of-fit test of a given distributions $F_0(x)$:

$$\boxed{\begin{array}{c} H_0 : F_X(x) = F_0(x) \quad \text{for all} \quad x \\ H_1 : F_X(x) \neq F_0(x) \quad \text{for some} \quad x \end{array}}$$

$$\boxed{\begin{array}{c} \text{Accept } H_0 \text{ is } D_n \leq D_{n\alpha} \\ \text{Reject } D_n > D_{n\alpha} \end{array}}$$

$$P_r\{D_n > D_{n\alpha}\} = \alpha \qquad D_n \sim ?$$

IE605:Engineering Statistics Manjesh K. Hanawal 16



KS one sided tests

$$D_n^+ = \sup_x S_n(x) - F_0(x)$$

$$D_n^- = \sup_x F_0(x) - S_n(x)$$

For α level test denote $D_{n\alpha}^+$ and $D_{n\alpha}^-$ denote $(1-\alpha)$th quartile of distribution of D_n^+ and D_n^- respectively.

$H_0 : F_X(x) \geq F_0(x)$ for all x
$H_1 : F_X(x) < F_0(x)$ for some x

Accept H_0 is $D_n^+ \leq D_{n\alpha}^+$
Reject $D_n^+ > D_{n\alpha}^+$

$H_0 : F_X(x) \geq F_0(x) \ \forall x$
$H_1 : F_X(x) < F_0(x)$ for some x

Accept H_0 if $D_n^- \leq D_{n\alpha}^-$
reject o.w.

IE605:Engineering Statistics Manjesh K. Hanawal 17

So, notice that when we use the Chi-square test, it basically compared f i with e i over k points, i equals to 1 to k. But we actually had n-points. So, then why to compare only this k points, why not compare all these n-points. So, that is what we will do here in the Kolmogorov-Smirnov test.

So, we are going to compare all the n-points and but these points will be compared at the empirical relative frequencies. So, what basically we are going to do is, we are going to compute the empirical, we are going to compute the deviation between the observed empirical cdf, and that CDF under the null hypothesis. So, we know that under the null hypothesis, my CDF is already given to me. And now I am going to compare it against my empirical CDF.

And when we want to apply this Kolmogorov-Smirnov test, we are going to assume that my null hypothesis corresponds to a continuous distribution. Now, let us try to understand what how to compute this empirical distribution. Suppose we have a random sample, then it is empirical CDF at point S is basically the fraction of the points that takes value less than or equals to X i.

So formally, this is basically our number of points which are taking value less than x divided by n. And in this case, in the, I am going to simply create as now KS test. The test statistics in KS test is the maximum deviation between my empirical and my expected CDFs. So, and x can be interval, so we are going to take x can be in fact, here is the entire real life. So, we are going to take the supremum over all these differences.

Now, the question here is does this D n? Statistics, if I want to find the statistics, do I need to know the underlying distribution of the sample? To compute the D n, yes, indeed, I need to know what is F not. But does distribution of the S D n itself depend on the underlying distribution. So, whatever it is, let us assume for hypothetically, whatever whether it depends or not, we will come to a point. But let us say whatever the D n we have, it is a stochastic quantity. And we can for a given alpha, let us denote D n alpha to be the 1 minus alpha quartile of the distribution.

Now, if you want use these statistics, to check my hypothesis, whether it has a CDF F not, which match at all the points, or it is going to be different at least one point. I can check this hypothesis by comparing my D n against this D n alpha. So, if I am going to accept this null hypothesis when D n is going to be small.

And as again if this and I am going to reject it if this D n is going to be larger than this D n alpha. So, this makes sense because if my empirical estimation, if my underlying samples are indeed following my null hypothesis distribution, then S n is expected to be close as to F not at most of the points. And if they it is going to follow something else, it is going to be deferring at least few points, or at least one point, and then this D n is going to be large.

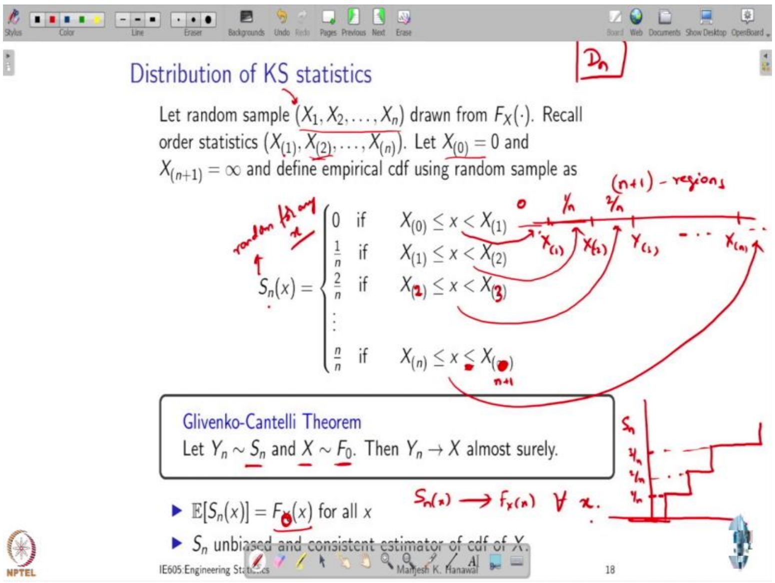
So, based on that intuition, we can set up this D n. But then by setting my D, my threshold here, which is D n alpha in this fashion, which is D n alpha is the 1 minus alpha quartile, I am going to get that the probability of reject to be alpha. That is, I am going to get an alpha level test. But does this how to compute the D n distribution? What is the distribution of D n? How is this D n distributed? We will come to that. So, before that this test we had you notice if you notice, we took the absolute difference. And when we wanted to check we wanted this to be exactly equal at all the points this is called a two-sided test of the KS test.

And when we do not take the absolute value, but we take the difference between S n and F 0, then we are going to define the statistics at D n plus, and when we take the difference between F 0 and S n, and take the maximal, all value of x, then we are going to give denoted D n minus.

And now if we are going to define alpha level tests by taking D n alpha plus and D n alpha minus to be 1 minus alpha-th quartile of the distributions, D n alpha and D n alpha minus respectively, then we can think of one-sided test. So, if you want to test that F x matches with F 0 only is going to be larger than F 0 at all the point, or it is going to be maybe this should be less than at least for some time, for some.

And we are going to accept or reject this hypothesis maybe when D n alpha D n plus is less than a D plus n alpha we are going to accept, and when we are if not the case, we are going to reject it. So, this we can do in this case. And when we want to test the opposite of this like whether my F of x is going to be for all x and or it is going to be less than F of x for some S, we can use the D n minus now, and then I am going to accept at 0 if D n minus is less than or equals to to the alpha minus, and similarly otherwise and reject otherwise. And these tests are called one sided test. So, and we have 2 criterias here depending on whether we want to see that my actual hypothesis is going to be always larger than minimal hypothesis at all the points, or it is going to be less.

(Refer Slide Time: 8:45)



Distribution of KS statistics

D_n

Let random sample $(X_1, X_2, \ldots, X_n)$ drawn from $F_X(\cdot)$. Recall order statistics $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$. Let $X_{(0)} = 0$ and $X_{(n+1)} = \infty$ and define empirical cdf using random sample as

$$S_n(x) = \begin{cases} 0 & \text{if} \quad X_{(0)} \le x < X_{(1)} \\ \frac{1}{n} & \text{if} \quad X_{(1)} \le x < X_{(2)} \\ \frac{2}{n} & \text{if} \quad X_{(2)} \le x < X_{(3)} \\ \vdots & \\ \frac{n}{n} & \text{if} \quad X_{(n)} \le x \le X_{(n+1)} \end{cases}$$

random for any x

$(n+1)$ - regions

0 $\quad 1/n \quad 2/n$

$X_{(1)} \quad X_{(2)} \quad Y_{(3)} \quad \ldots \quad X_{(n)}$

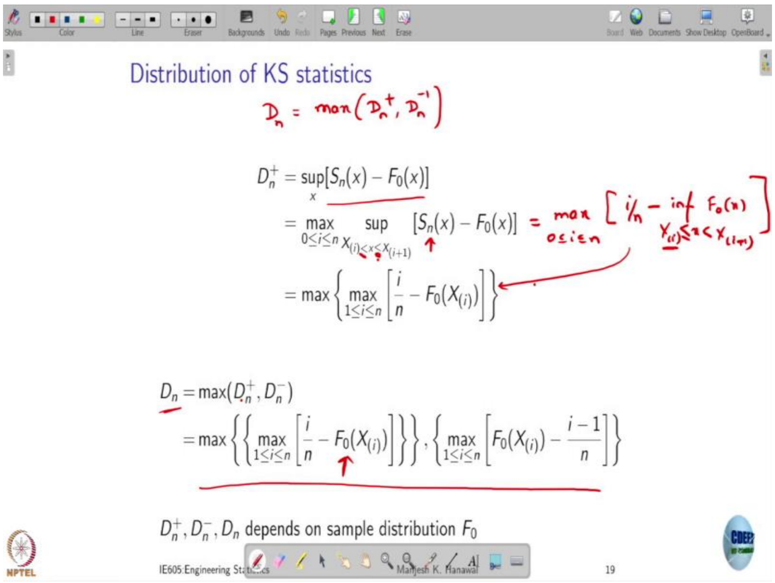
Glivenko-Cantelli Theorem

Let $Y_n \sim S_n$ and $X \sim F_0$. Then $Y_n \to X$ almost surely.

- $\mathbb{E}[S_n(x)] = F_X(x)$ for all x
- S_n unbiased and consistent estimator of cdf of X.

$S_n(x) \longrightarrow F_X(x) \ \forall \ x.$

IE605:Engineering Statistics Manjesh K. Hanawal 18



Distribution of KS statistics

$D_n = \max(D_n^+, D_n^-)$

$$D_n^+ = \sup_x [S_n(x) - F_0(x)]$$

$$= \max_{0 \le i \le n} \sup_{X_{(i)} \le x \le X_{(i+1)}} [S_n(x) - F_0(x)] = \max_{0 \le i \le n} \left[i/n - \inf_{X_{(i)} \le x < X_{(i+1)}} F_0(x) \right]$$

$$= \max \left\{ \max_{1 \le i \le n} \left[\frac{i}{n} - F_0(X_{(i)}) \right] \right\}$$

$$D_n = \max(D_n^+, D_n^-)$$

$$= \max \left\{ \left\{ \max_{1 \le i \le n} \left[\frac{i}{n} - F_0(X_{(i)}) \right] \right\} \right\}, \left\{ \max_{1 \le i \le n} \left[F_0(X_{(i)}) - \frac{i-1}{n} \right] \right\}$$

D_n^+, D_n^-, D_n depends on sample distribution F_0

IE605:Engineering Statistics Manjesh K. Hanawal 19

To understand now, the distribution of D n, we need to little bit revisit the properties of empirical distribution and also our order statistics. Let us say I have a random sample drawn from some underlying distribution. Now, recall that we have denoted the order statistics as X of 1 until this parenthesis like this and X of 2 and 2 onto this parenthesis like this, where X 1 denoted the smallest value and this quantity denoted the second smallest value like that. Now, instead of only looking at the first order and nth order statistics, we can also look into the 0th order statistics, but it is just simply defined to be 0. And also, we can define n plus 1th order statistics and which will be simply take it as infinity.

Now, with this, we can write our empirical distribution as this, it is going to be 0 when X is going to be, maybe let me write this. I have X of 1 here, X of 2 here and a X of 3 here like that, and I have X of n like this. Then so, all this region is what this entire region here is captured by this, and this region here is captured by this, and I think I made a mistake here this should be 2, and 3 and this region is captured by this, and 3 and with this infinite is not there and this and this should have been n plus 1 here, and this region is captured by this. Now, it so happens that, if I have a sequence of random variables Y n, where Y n is going to be distributed as S n and I have a random variable S, X, which is distributed as per my null hypothesis, then one can argue that Y n converges to X almost surely.

Further if I take, so, notice that this S n is a random quantity for any X because this itself is defined in terms of this random samples. Now, it so happens that the expected value of S n at point X is simply F of X, or maybe I should have written this F of 0 computed at S. So, what we are basically saying is and also, we can show that basically we can show that this is going to be F of X, X for all X. So, then what we are saying is S n is an unbiased and consistent estimator of the CDF of x. So, this S n is going to provide me a good information about my null hypothesis when I have a large number of samples. So, that is why I want to compare these two.

Now to understand the distribution let us little bit express this D n and use the properties of my order statistics. So, I know that D n can be written as D n plus and D n minus. Now, D n plus is this simply and I can split this summation over all into max over my i’s, between 0 to n, and also between the ranges.

So, maybe I think I made a mistake here, this should have been X minus so basically, I am the soup. These super entire regions, I am basically dividing into this region. So, this entire region I have basically divided into n plus 1 regions, and I am looking now at the max in each of these regions.

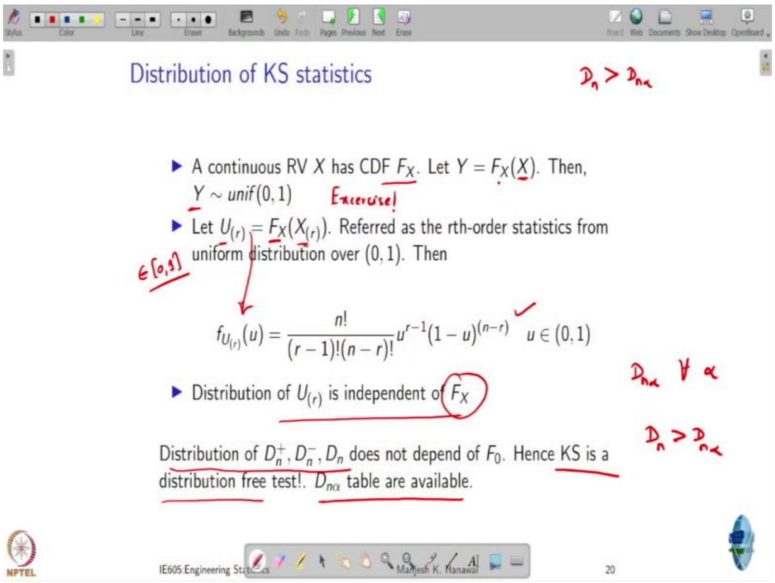
Now, I know that for a X, which between X of i and X i plus 1, my S n is going to be constant, and that is given by i by n. And so, what I mean here is like, we just said that this is going to be the value of my S n is 0 here, 1 by n in this range, 2 by n in this range like that. So, we know that my S n is like maybe like somewhere like it jumps like this.

Maybe wherever it is like it jumps like this by equal number of amount. And maybe like this, so this is like 1 by n, this is like 2 by n, and this is like the 3 by n wherever it is. So now, if I know that if I max over 0 into n, I know that this is if I take soup inside because this guy I know is simply i by n in that range. And this is now going to be in for X of i less than or equal to X, less than X of i plus 1 of F 0 X.

And I know that the my F 0 because of its monotonicity properties and I know that and if I have to take its infimum value, it is going to be the smallest here. And that is why I am going to when I go from here this info over this I can simply write F of 0 X of i. Now, one can also compute D n minus similarly, using D n minus and D n plus I can write this expression like this.

But notice that even after writing this, I have just did give a little longer expression, but both D n plus, D n minus, D n here all of them are depending on this F not, which is basically the distribution of my underlying like off my hypothesis, my null hypothesis. At this point, it is not clear. Why is that this D n has distribution is independent of these underlying distributions for which we are testing the samples against.

(Refer Slide Time: 16:49)



Distribution of KS statistics

$D_n > D_{n\alpha}$

- A continuous RV X has CDF F_X. Let $Y = F_X(X)$. Then, $Y \sim unif(0,1)$ Excercise!
- Let $U_{(r)} = F_X(X_{(r)})$. Referred as the rth-order statistics from uniform distribution over $(0,1)$. Then

$\in [0,1]$

$$f_{U_{(r)}}(u) = \frac{n!}{(r-1)!(n-r)!} u^{r-1}(1-u)^{(n-r)} \quad u \in (0,1)$$

✓

- Distribution of $U_{(r)}$ is independent of F_X

$D_{n\alpha} \ \forall \alpha$

Distribution of D_n^+, D_n^-, D_n does not depend of F_0. Hence KS is a distribution free test!. $D_{n\alpha}$ table are available.

$D_n > D_{n\alpha}$

To do this further, let us understand, or let us understand little bit more of the properties and the transformations of the random variable. Suppose I have a random variable X, which has a CDF of F of X. And if I define a new random variable by applying transformation F of X on X, then my new random variable Y has uniform distribution.

So, you can check this is like an exercise. Now, I can define a new random variable by applying this transformation F of X on the rth order statistics. Instead of simply taking my random variable X, I am going to now declare this X by its rth order sample, and I am going to define that new value as U r. Now, this is going to refer to as rth order statistics from uniform distribution over 0, 1.

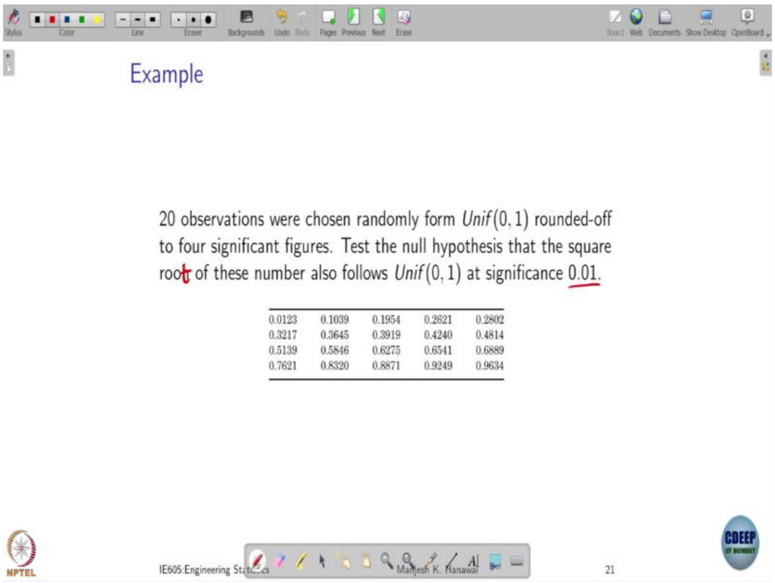
So, notice that it is range U of r, it is again going to be between 0, 1. And we are going to call it as rth order statistic from uniform distribution, and one can explicitly derive the distribution of this F U of r and this is given like this. Now, if you notice this, this distribution of U r is independent of F of X, and this is I think this is like a beta distribution.

And it does not depend on underlying CDF, which we started with, which has like whatever the F of X we started with, because of that, the distribution of the statistics we were interested in does not depend on F not. And one can compute, even though its explicit form of this distributions is not available. One can do numerical computations, and get the tail probabilities of these distributions and one can compute D n of alpha for all values of alpha.

And now we can apply these thresholds and to see whether I want to get a alpha level test. And because of that, we can conclude that the KS test is a distribution free test, because it does the your statistics, you do not need to make any assumption about that, it is independent of what is the null hypothesis that you want to interested in, unlike in the t test or F test, where we have to explicitly assume that your statistics is going to be either T distributed or F distribution. So, that is why we are going to call it as a distribution test.

And here as I said, because this D n alpha does not depend on any test, one can do extensive numerical simulations and get a very good approx, very good values of this tail probabilities of this D n. So, now, I hope, how to apply KS test is clear, all you need to do is compute your D n, if you want to do a 2-sided test, and if you want to get a alpha level test, you compare it against D alpha and if this is larger, you reject it. And this is going to give you a alpha level test.

(Refer Slide Time: 21:09)

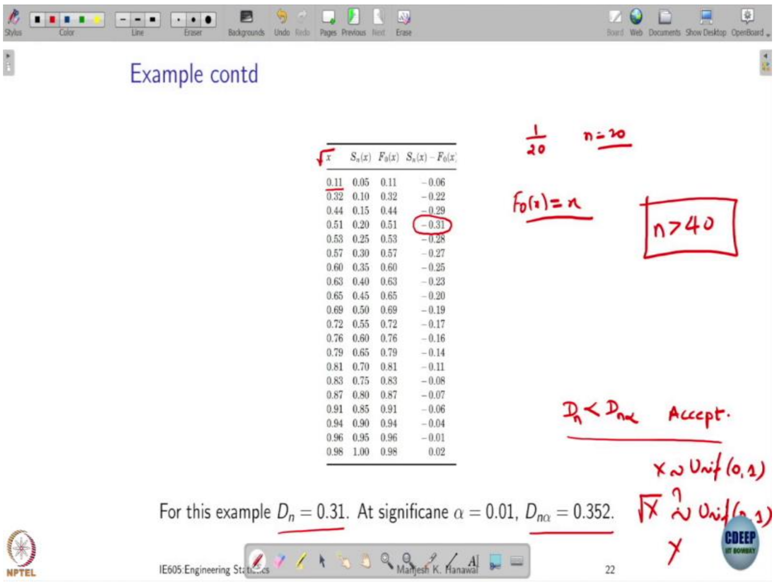


Example

20 observations were chosen randomly form $Unif(0,1)$ rounded-off to four significant figures. Test the null hypothesis that the square root of these number also follows $Unif(0,1)$ at significance 0.01.

0.0123	0.1039	0.1954	0.2621	0.2802
0.3217	0.3645	0.3919	0.4240	0.4814
0.5139	0.5846	0.6275	0.6541	0.6889
0.7621	0.8320	0.8871	0.9249	0.9634

IE605:Engineering Statistics Manjesh K. Hanawal 21



Example contd

$\sqrt{x}$	$S_n(x)$	$F_0(x)$	$S_n(x) - F_0(x)$
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

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

$F_0(x)=x$

$\boxed{n>40}$

$D_n < D_{n\alpha}$ Accept.

$X \sim Unif(0,1)$

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

X

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

IE605:Engineering Statistics Manjesh K. Hanawal 22

As s quick example, suppose let us say there are 20 observations were chosen uniformly random, over 0, 1 interval, and they were rounded up to 4 significant digits. And you want to test the null hypothesis that the square root of these numbers follow uniform distribution again or interval 0 at significance 0.1.

That 20 samples are taken and they are organised in increasing order here. And now, let us see how we can apply KS test here. So, now that table it is the computations are represented in the table format here. So, because we are interested in the square root of the observed sample, so we take the square root of this values here, and their square roots are written here, this is basically actually the square root of X now, like let us say these are X here, and this is like a square root of X here.

And we know that S n is going to increment in value of 1 by 20 here because n equals to 20 here, so you will see that these values are increasing in values of 0.05. And this being a uniform distribution, the null hypothesis being uniform distribution, we know that this value is going to be same as this value, because for a uniform distribution, we know this is equals to F of not x equals to simply X, or this like a linear line.

And now you can take the difference between them. And what we will be interested in in the absolute value of this which is the maximum, if you look into this absolute value, this is the one which has the highest value and that is going to the value of D n, at significance value alpha goes to 0.1 from the table so you can get it to be 0.352.

And now you will see that D n is going to be less than a D, than alpha. So, because of that, you, your test suggest that you are going to accept. But then is this correct here. So, notice that what we have been told is X is uniform 0, 1 And we have been asked to check square root of X is also uniform, which is not the case.

But in this case by applying this KS test, we ended up accepting that the square root of X is also a uniform distribution. So obviously, this is not correct. And here this number of samples are not good enough to make a decision. And usually as a thumb rule, one needs more than 40 number of samples to have a fairly correct answer, otherwise, there may be wrong conclusions will end up making.