One Sample KS-Test

Kolmogorov-Smirnov test is a non-parametric procedure used to test whether sample observations selected are drawn from specified theoretical population. It is an alternative to chi-square test for goodness of fit between observed and expected frequencies.

The null and alternative hypotheses of the test are:

H_0 : samples drawn are from population having specified (empirical) theoretical distribution (population)

H_0 : samples drawn are not from population having specified (empirical) theoretical distribution (population)

For the test procedure first given observations $X_1, X_2, \ldots \ldots \ldots \ldots., X_n$ are arranged in ascending order to get order statistic $y_1, y_2, \ldots \ldots \ldots, y_k$. If some of the observations are repeated, then corresponding frequencies are obtained as $f_1, f_2, \ldots \ldots \ldots, f_k$and then cumulated frequencies are calculated as $n_1, n_2, \ldots \ldots \ldots.., n_k$.

Next, empirical distribution function is calculated by formula

#.(a) for observations which are not repeated

$$F_e(y) = \frac{k}{n}, \ for\ y_k \le y < y_{k+1}, \qquad k = 1, 2, \ldots \ldots \ldots \ldots, n-1$$

#.(b) for observations with repeated values

$$F_e(y) = \frac{n_k}{n}, \ for\ y_k \le y < y_{k+1}, \qquad k = 1, 2, \ldots \ldots \ldots \ldots, n-1$$

Next, according to null hypothesis that samples drawn are from population having specified theoretical distribution (population), the hypothetical distribution function $F_o(y)$ (which is distribution function if null hypothesis H_0 is true) is obtained.

(The way of finding $F_0(y)$ depends on population distribution under consideration and it is described in different problems below.)

Then following two absolute differences between $F_e(y) and\ F_o(y)$ are calculated:

$$|F_0(y) - F_e(y)|$$

$$|F_0(y-1) - F_e(y)|$$

Let

$$D_1 = \max|F_0(y) - F_e(y)|$$

$$D_2 = \max|F_0(y-1) - F_e(y)|$$

Then following value is considered as the test statistic

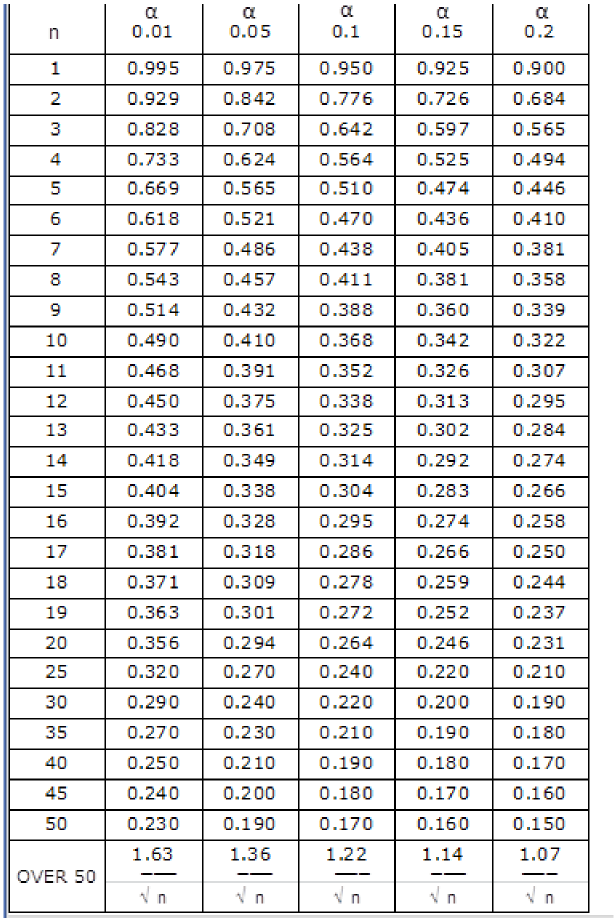
$$D_0 = \max(D_1, D_2)$$

Critical value for rejecting H$_0$ at α level of significance and for given value of 'n', denoted as D$_{\alpha,n}$ is obtained as follows: (i) for n <= 40, we observe value in Kolmogorov-Smirnov table for prescribed value of α and for given value of n (given below), (ii) for n > 40, it is given by

α =	0.01	0.05	0.1	0.15	0.2
$D_{\alpha,n}$=	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$

Finally H$_0$ is rejected if $D_0 \geq D_{\alpha,n}$.

KS-Table



n	α 0.01	α 0.05	α 0.1	α 0.15	α 0.2
1	0.995	0.975	0.950	0.925	0.900
2	0.929	0.842	0.776	0.726	0.684
3	0.828	0.708	0.642	0.597	0.565
4	0.733	0.624	0.564	0.525	0.494
5	0.669	0.565	0.510	0.474	0.446
6	0.618	0.521	0.470	0.436	0.410
7	0.577	0.486	0.438	0.405	0.381
8	0.543	0.457	0.411	0.381	0.358
9	0.514	0.432	0.388	0.360	0.339
10	0.490	0.410	0.368	0.342	0.322
11	0.468	0.391	0.352	0.326	0.307
12	0.450	0.375	0.338	0.313	0.295
13	0.433	0.361	0.325	0.302	0.284
14	0.418	0.349	0.314	0.292	0.274
15	0.404	0.338	0.304	0.283	0.266
16	0.392	0.328	0.295	0.274	0.258
17	0.381	0.318	0.286	0.266	0.250
18	0.371	0.309	0.278	0.259	0.244
19	0.363	0.301	0.272	0.252	0.237
20	0.356	0.294	0.264	0.246	0.231
25	0.320	0.270	0.240	0.220	0.210
30	0.290	0.240	0.220	0.200	0.190
35	0.270	0.230	0.210	0.190	0.180
40	0.250	0.210	0.190	0.180	0.170
45	0.240	0.200	0.180	0.170	0.160
50	0.230	0.190	0.170	0.160	0.150
OVER 50	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$

Problem:

#.1

Can following random numbers generated by a pRNG be considered to be uniformly distributed? Use Kolmogorov-Smirnov test at 5% level of significance. The random numbers are:

0.85245902 0.75409836 0.59016393 0.98360656 0.63934426

Solution-

Here ordered statistics is

0.59016393 0.63934426 0.75409836 0.85245902 0.98360656

Here, k = 5, n = 5.

Working Table-

k	y	$F_e(y) = \frac{k}{n}$	$F_0(y) = \frac{k}{count\ of\ k}$	$\lvert F_e(y) - F_0(y)\rvert$	$F_0(y-1)$	$\lvert F_e(y) - F_0(y-1)\rvert$
1	0.590164	0.2	0.2	0	0	0.2
2	0.639344	0.4	0.4	0	0.2	0.2
3	0.754098	0.6	0.6	0	0.4	0.2
4	0.852459	0.8	0.8	0	0.6	0.2
5	0.983607	1	1	0	0.8	0.2
n = 5				$D_1 = 0$		$D_1 = 0.2$

Test statistic is

$$D_0 = \max(D_1, D_2) = 0.2$$

From KS-table for $\alpha = 0.05\ and\ n = 5, we\ have\ D_{\alpha,n} = D_{0.05,5} = 0.565$

Since it is not true that $D_0 > 0.565$, thus the null hypothesis is not rejected and it is concluded that ....

#.2

A die is rolled 60 times and following values were obtained as a sequence of random numbers: 1, 2, 1, 3, 3, 4, 1, ... The frequency distribution of number of values obtained is given below-

Side	1	2	3	4	5	6
# of times observed	8	9	13	7	15	8

Using Kolmogorov-Smirnov method to test whether the values obtained in throwing of dice are uniform? Test at 5% level of significance.

Solution-

Here the null and alternative hypothesis are:

$$\boldsymbol{H_0: \ the\ values\ in\ throwing\ of\ die\ are\ uniform.}$$

$$\boldsymbol{H_1: \ the\ values\ in\ throwing\ of\ die\ are\ not\ uniform.}$$

Here, k = 6 and n = 60.

Working Table-

k	y	$freq.$ (f)	$c.f.$ (n_k)	$F_e(y) = \frac{n_k}{n}$	$F_0(y) = \frac{k}{Count\ of\ k}$	$\lvert F_e(y) - F_0(y)\rvert$	$F_0(y-1)$	$\lvert F_e(y) - F_0(y-1)\rvert$
1	1	8	8	0.133	0.167	0.033	0.000	0.133
2	2	9	17	0.283	0.333	0.050	0.167	0.117
3	3	13	30	0.500	0.500	0.000	0.333	0.167
4	4	7	37	0.617	0.667	0.050	0.500	0.117
5	5	15	52	0.867	0.833	0.033	0.667	0.200
6	6	8	60	1.000	1.000	0.000	0.833	0.167
		n=60				$D_1 = 0.050$		$D_2 = 0.200$

Test statistic is

$$D_0 = \max(D_1, D_2) = \max(0.05, 0.2) = 0.2$$

Critical value at α = 0.05 is $D_{0.05,60} = \frac{1.36}{\sqrt{60}} = 0.176$

Since $D_0 > 0.176$, so H_0 is rejected.

#.3

Frequency distribution of 10 photographs captured by a camera with respect to its brightness is given below:

Brightness	Very Dark	Dark	Normal	Bright	Very Bright
Frequency	0	1	0	5	4

Use KS-test to observe whether the number of photographs of different brightness are uniformly generated random numbers.

Solution-

Here,

H_0: the number of photographs of different brightness are uniformly generated random numbers

H_1: the number of photographs of different brightness are not uniformly generated random numbers

Working table:

k	y	$freq.$ (f)	$c.f.$ (n_k)	$F_e(y) = \frac{n_k}{n}$	$F_0(y) = \frac{k}{Count\ of\ k}$	$\lvert F_e(y) - F_0(y)\rvert$	$F_0(y-1)$	$\lvert F_e(y) - F_0(y-1)\rvert$
1	v.d.	0	0	0	0.2	0.2	0	0
2	d.	1	1	0.1	0.4	0.3	0.2	0.1
3	n.	0	1	0.1	0.6	0.5	0.4	0.3
4	b.	5	6	0.6	0.8	0.2	0.6	0
5	v.b.	4	10	1	1	0	0.8	0.2
		n=10				$D_1 = 0.5$		$D_2 = 0.3$

The test statistic is

$$D_0 = \max(D_1, D_2) = \max(0.5, 0.3) = 0.5$$

Critical value at $\alpha = 0.05$ is $D_{0.01,10} = 0.490$

Since $D_0 > 0.49$, so H_0 is rejeted.

#.4

An emergency ward of a newly constructed hospital in a village consists of 20 beds. An analyst in due course of study in between 2 pm to 4 pm found the distribution of bed occupants during different seasons is as follows:

Occupants (x)	Summer	Fall	Winter	Spring
Frequency (f)	8	5	20	15

Do the data provide sufficient evidence to indicate that the number of bed occupants in the hospital are uniformly generated random numbers? Use K-S test at 0.05 level of significance.

Solution-

Here

H$_0$: number of bed occupants in the hospital are uniformly generated random numbers

H$_1$: number of bed occupants in the hospital are not uniformly generated random numbers

Working Table:

k	y	$freq.$ (f)	$c.f.$ (n_k)	$F_e(y) = \frac{n_k}{n}$	$F_0(y) = \frac{k}{Count\ of\ k}$	$\lvert F_e(y) - F_0(y)\rvert$	$F_0(y-1)$	$\lvert F_e(y) - F_0(y-1)\rvert$
1	su	8	8	0.167	0.250	0.083	0.000	0.167
2	f	5	13	0.271	0.500	0.229	0.250	0.021
3	w	20	33	0.688	0.750	0.063	0.500	0.188
4	sp	15	48	1.000	1.000	0.000	0.750	0.250
		n=48				$D_1 = 0.229$		$D_2 = 0.250$

The test statistic is

$$D_0 = \max(D_1, D_2) = \max(0.229, 0.250) = 0.250$$

Critical value at α = 0.05 is $D_{0.05,48} = \frac{1.36}{\sqrt{48}} = 0.196$

Since $D_0 > 0.196$, so H$_0$ is rejected.

Two- Sample KS-Test:

KS test can also be used to test whether two independent samples under consideration have been drawn from the similar populations or from the same population.

The null and alternative hypotheses of the test are:

H_0: two independent samples have been drawn from the same or identical population.

H_1: the population from which first sample is drawn have less r greater of different value of study characteristic than the population from which second sample is drawn.

Accordingly it may be right tailed or left tailed or two tailed.

The test procedure for determining test statistic is same as that of one sample KS-test, except that instead of calculating difference between observed cdf and empirical cdf, as measured in one-sample test, we calculate absolute value of difference in cdf of first and second sample in two sample KS-test, i.e., $|F_1(x) - F_2(x)|$, where F_1(x) and F_2(x) are observed cdf's obtained from first sample and second sample respectively.

Let n_1 and n_2 denote total frequencies of first and second samples.

Case (a) When $n_1 = n_2$

In this case the test statistic is given by:

$$D_0 = \max|F_1(x) - F_2(x)|$$

In case when $n_1 = n_2$, then process of determining critical value is same as that of one-sample KS-test.

Case (b) – when $n_1 \neq n_2$

In this case after calculating

$$D_0 = \max|F_1(x) - F_2(x)|$$

the value of following test statistic is calculated:

$$\chi^2 = \frac{4\,D_0^2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

which has Chi-square distribution with degrees of freedom given by

$$\upsilon = \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Finally H0 is rejected at α level if (i) for right-tailed test $\chi^2 \geq \chi^2_{\alpha,\upsilon}$ (ii) $\chi^2 \leq \chi^2_{1-\alpha,\upsilon}$ (iii) for two-tailed test $\chi^2 \geq \chi^2_{\alpha/2,\upsilon}$ or $\chi^2 \leq \chi^2_{1-\alpha/2,\upsilon}$

Problem:

#.1

(for $n_1 = n_2$)

In order to ascertain whether there is any significant difference in the monthly earnings of teachings of technicians and administrative, a random sample of 30 technicians and 30 administrative was selected and their earnings recorded. The following table gives the distribution of monthly earnings of technicians and administrative:

Monthly Earnings (Rs.)	Number of Persons	
	Technicians	Administrative
2000-4000	1	4
4000-6000	3	5
6000-8000	6	8
8000-10000	9	7
10000-12000	5	4
12000-14000	4	2
14000-16000	2	0

Use Kolmogorov-Smirnov two sample test to find if there is any significant difference in the earnings of technicians and administrative.

Solution-

Here,

H_0: there is no significant difference between earnings of technicians and administrative

H_1: there is significant difference between earnings of technicians and administrative

	Number of Persons		Probabilities		cdfs		$\lvert F_1(x)- F_2(x)\rvert$
	Technicians (f_1)	Administrative (f_2)	$p_1(x)$	$p_2(x)$	$F_1(x)$	$F_2(x)$	
2000-4000	1	4	1/30	4/30	1/30	4/30	3/30
4000-6000	3	5	3/30	5/30	4/30	9/30	5/30
6000-8000	6	8	6/30	8/30	10/30	17/30	7/30
8000-10000	9	7	9/30	7/30	19/30	24/30	5/30
10000-12000	5	4	5/30	4/30	24/30	28/30	4/30
12000-14000	4	2	4/30	2/30	28/30	30/30	2/30
14000-16000	2	0	2/30	0/30	30/30	30/30	0/30
Total	30	30					

Test statistic is $D_0 = \max(|F_1(x) - F_2(x)|) = 7/30 = 0.233$

Critical value at $\alpha = 0.05$ is $D_{0.05,30} = 0.2417$

Since $D_0 \ngeq 0.2471$, so H_0 is not rejected.

#.2

(for $n_1 \neq n_2$)

In Russia, political prisoners are exiled to either Siberia or to the Urals where they are destined to die within short period of time due to the effect of radiation. An analyst seeking the life expectancy of the exiled prisoners found the following data:

Survive Yrs.	0-1	1-2	2-3	3-4	4+
Siberia	8	12	16	10	4
The Urals	5	7	31	12	5

Test at 5% level of significance if the difference of survival time is significantly different by using Kolmogorov-Smirnov method.

Solution-

Here,

H_0: there is no significant difference between survival periods of two places

H_1: there is significant difference between survival periods of two places

Working Table-

Survival Yrs	Number of Deaths		Probabilities		cdfs		$\lvert F_1(x) - F_2(x)\rvert$
	Siberia (f_1)	The Urals(f_2)	$p_1(x)$	$p_2(x)$	$F_1(x)$	$F_2(x)$	
0-1	8	5	8/50	5/60	8/50	5/60	0.077
1-2	12	7	12/50	7/60	20/50	12/60	0.200
2-3	16	31	16/50	31/60	36/50	43/60	0.003
3-4	10	12	10/50	12/60	46/50	55/60	0.003
4+	4	5	4/50	5/60	50/50	60/60	0.000
Total	50	60					

Now, $D_0 = \max|F_1(x) - F_2(x)| = 0.200.$

Here n1 = 50 and n2 = 60, since $n_1 \neq n_2$, so test statistic is

$$\chi^2 = \frac{4\,D_0^2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{4x\;0.200^2}{\frac{1}{50} + \frac{1}{60}} = \frac{0.16}{0.037} = 4.36$$

The degrees of freedom of above statistic is

$$v = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \sqrt{\frac{50x60}{50 + 60}} = 5.22 = 5 \ (rounded)$$

It is two tailed test, so critical values are ${\chi^2}_{\alpha/2,v} = {\chi^2}_{0.025,5} = 12.833$ and ${\chi^2}_{1-\alpha/2,v} = {\chi^2}_{0.975,5} = 0.831.$

Since neither $\chi^2 \geq 12.833$ nor $\chi^2 \leq 0.831$, so H_0 is not rejected.