Kolmogorov-Smirnov Goodness-of-fit Tests

October 16, 2023

This test is used in situations where a comparison has to be made between an observed sample distribution and a theoretical distribution. The *goodness-of-fit test* that we will learn about about was developed by two probabilists.

Andrey Kolmogorov(1903-1987) and Nikolai Smirnov(1900-1966)

In this, we will learn how to conduct a test to see how well a hypothesized distribution function $F(x)$ fits an empirical distribution function $F_n(X)$.

Empirical distribution function

Given an observed random sample $X_1, X_2, \ldots, X_n$ an empirical distribution function $F_n(x)$ is the fraction of sample observations less than or equal to the value x.

More specifically, if $x_1 < x_2 < \ldots, < x_n$ are the order statistics of the observed random sample, with no two observations being equal, than the empirical distribution function is defined as

$$F_n(x) = \begin{cases} 0; & x < x_1, \\ \frac{k}{n}; & x_k \le x \le x_{k+1},\ k = 1, 2, \ldots, n-1 \\ 1; & x \ge x_n \end{cases}$$

Example: A random sample of 8 people yields the following counts the number of times they read in the past months;

$$2, 1, 0, 2, 6, 6, 4, 7$$

Calculate the empirical distribution function $F_n(x)$.
Solution: The empirical distribution function is

sample	x	0	1	2	4	6	7	
Frequency	f	1	1	2	1	2	1	
PDF	$f_X(x)$	1/8	1/8	2/8	1/8	2/8	1/8	
CDF	$F_X(x)$	1/8	2/8	4/8	5/8	7/8	8/8	$\rightarrow F_n(x)$; empirical distribution

Results: *Let X be continuous random variable with a CDF F and let $U = F(X)$. Then $U \sim U[0, 1]$.*

K-S approximation to Null Distribution:

Kolmogorov-Smirnov (K-S) tests statistics D_n is defined as

$$D_n = \sup |F_0(x) - F(x)|, \ x \in \mathbb{R},$$

where $F_0(x) =$Theoretical distribution; $F(x) =$Empirical distribution and $n =$total number of data sets.
Results: *Let F be a continuous CDF and let $X_1, X_2, \ldots, X_n$ be a sequence of iid rv's with CDF F. Then*

1. *The null distribution of D_n does not depends on F_0 (Non-parametric distribution); it depends only on n.*

2. *If $n \to \infty$, the distribution of $\sqrt{n}D_n$ is asymptotically Kolmogorov-Smirnov's distribution with the CDF*

$$Q(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}$$

That is,

$$\lim n \to \infty P(\sqrt{n}D_n \leq x) = Q(x).$$

K-S One Sample Test:

- Small sample: when sample size $n < 30$.
- Large sample: when size size $n \geq 30$.

Hypothesis (3 step rule)

I) Define the Hypothesis:

- Null Hypothesis:
$$H_0: \ F_0(x) = F_n(x)$$
i.e. assume no difference between the observed (empirical) $F_n(x)$ and theoretical $F_0(x)$ distribution.
- Alternative Hypothesis:
$$H_1: \ F_0(x) \neq F_n(x)$$

II) Compute test statistics:

- Order sample x as $(x_1 < x_2 < \cdots < x_n)$.
- Value of test statistics $D = \sup |F_0(x) - F_n(x)|$.
Note: *The rule for computing the KS test statistic:*
For each ordered observation, x_k computes the differences
$$\begin{aligned} D^+ &= \max_{1 \leq k \leq n} |F_0(x_k) - F_n(x_k)| \\ &= \max \left| F_0(x_k) - \frac{k}{n} \right|, \ k = 1, 2, \ldots, n \end{aligned}$$
and
$$\begin{aligned} D^- &= \max_{1 \leq k \leq n} |F_0(x_k) - F_n(x_{k-1})| \\ &= \max \left| F_0(x_k) - \frac{k-1}{n} \right|. \end{aligned}$$
The largest of these is the K-S test statistic and the test statistics value is
$$D = \max(D^+, D^-)$$

III) Conclusion:
The critical value of D is found from the K-S table values for one sample test.

- Reject H_0 : if $D >$ critical value.
- Accepted H_0 : if $D <$ critical value.

Example: Consider the data points $1.1, 0.26, 1.97, 0.33, 0.55, 0.77, 1.46, 1.18$; Is there any evidence to suggest that the data were not randomly samples from a Uniform$(0, 2)$ distributed?

Solution. Step I:

Define the hypothesis:

$$H_0 : F(x) = F_0$$
$$H_1 : F(x) \neq F_0$$

Where $F(x)$ is the unknown CDF from which our data were sampled, and $F_0(x)$ is the CDF of the Uniform(0,2) distribution.

The PDF of Uniform(0,2) is

$$f(x) = \frac{1}{2};\ 0 < x < 2$$

and CDF

$$F_0(x) = P(X \leq x) = \begin{cases} 0; & 0 < x \\ \frac{x}{2}; & 0 < x < 2 \\ 1; & x \geq 2 \end{cases}$$

And the empirical CDF $F(x)$ satisfies

$$F_n(x_n) = \frac{k}{8} \text{ for } k = 1, 2 \ldots, 8. \text{ } (\because \text{ the number of samples is 8.})$$

Step II:

Compute observed and theoretical distribution.

k	x_k	$F_n(x_k)$	$F_0(x_k)$	$D^+ = \lvert F_0(x_k) - F_n(x_k)\rvert$	$F_n(x_{k-1})$	$D^- = \lvert F_0(x_k) - F_n(x_{k-1})\rvert$
1	0.26	$\frac{1}{8} = 0.125$	$\frac{0.26}{2} = 0.130$	0.005	0.000	**0.130**
2	0.33	$\frac{2}{8} = 0.250$	$\frac{0.33}{2} = 0.165$	0.085	0.125	0.040
3	0.55	0.357	0.275	0.100	0.250	0.025
4	0.77	0.500	0.385	0.115	0.375	0.010
5	1.18	0.625	0.590	0.035	0.500	0.090
6	1.41	0.750	0.705	0.045	0.625	0.080
7	1.46	0.875	0.730	**0.145**	0.750	0.020
8	1.97	$\frac{8}{8} = 1$	0.985	0.015	0.875	0.090

The test statistics D is

$$D = \max \lvert F_0(x) - F_n(x)\rvert = 0.145$$

OR

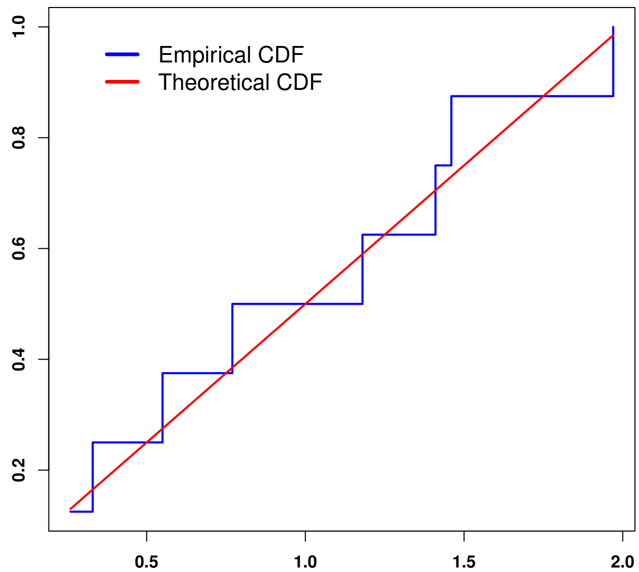
$$D = \max(D^+, D^-) = \max(0.130, 0.145) = 0.145$$

Setp III:

Since $n = 8 < 30$. So, we apply small sample K-S test. From the K-S table, the critical value of D is 5% significance level with $n = 8$ is given by

$$D_{0.05} = 0.454$$

As $0.145 < 0.454$, Hence we fail to reject the null hypothesis and conclude that the were sampled from Uniform(0,2) distributed.



Kolmogorov-Smirnov Test
Critical Values

SAMPLE SIZE (N)	LEVEL OF SIGNIFICANCE FOR D = MAXIMUM [$F_0(X)$ - $S_n(X)$]				
	.20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.210	.220	.240	.270	.320
30	.190	.200	.220	.240	.290
35	.180	.190	.210	.230	.270
OVER 35	$\frac{1.07}{\sqrt{N}}$	$\frac{1.14}{\sqrt{N}}$	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$