Kolmogorov Smirnov test (Single sample case)

Function of the test: The Kolmogorov Smirnov test is also called the K-S D test and it is nonparametric test. It is basically goodness of fit test. In one sample situation, K-S test determines how well a hypothesized frequency distribution $F_T(x)$ fits to observed (empirical) frequency distribution $F_O(x)$. It is more powerful alternative to chi-square of goodness of fit test. It does not require that numerical data to be divided into categories. It is more stringent test.

KS test in one sample case used for two situations:

1. Kolmogorov Smirnov goodness of fit for discrete for grouped data
2. Kolmogorov Smirnov goodness of fit for continuous data

Test Assumptions

1. Sample is drawn randomly from population
2. Hypothesized distribution is discrete or continuous, but the preferred distribution is continuous.
3. Hypothesized distribution is specified in advance. For example, if the hypothesized distribution is normal probability distribution, then the expected mean and standard deviation must be specified in advance.

Hypothesis

H_0: The data follow the specified distribution or there is no significant difference between observed frequency distribution and the distribution specified i.e. $PDF_T = PDF_O$

H_1: The data do not follow the specified distribution or there is significant difference between observed frequency distribution and the distribution specified i.e. $PDF_T \neq PDF_O$

In K-S test a comparison is made between some theoretical cumulative distribution function $F_T\left(x\right)$ and observed cumulative distribution function $F_O\left(x\right)$.

Test Statistic

The difference between the theoretical cumulative distribution function F_T (x) and the observed or empirical cumulative distribution function F_O (x) is measured by D statistic. To find the D statistic we have calculate two differences for each category or value.

$D_i = \left|F_O(x_i) - F_T(x_i)\right|$ and

$D_i^{/} = |F_O(x_{i-1}) - F_T(x_i)|$ for each i.

Then D is the largest D_i or the largest $D_i^{/}$ whichever is larger.

D = Max {Max D_i, Max $D_i^{/}$ }

The value of D may also be calculated graphically by actually measuring the largest vertical distance between the curves of $F_O(x)$ and $F_T(x)$.

For the discrete distribution, the test statistic is given by,

$$D_i = Max\left|F_O(x_i) - F_T(x_i)\right| \text{ for each i}$$

Decision Rule

The K-S test is right sided test. We will adopt following decision rule for continuous distribution.

Reject H_0 if Calculated D ≥ Critical D for n sample size.

Limitation of KS test

1. The distribution must be fully specified. K-S test is not appropriated when the parameters have to be estimated from the sample. If one or more parameters have to be estimated from the sample data, the test becomes conservative.
2. It is appropriate when the distribution is continuous. When D values are based on a discrete theoretical distribution the test becomes conservative.
3. K-S test is also conservative if continuous data are grouped.

Numerical Example:

Single sample case, Discrete data

A dice is rolled up for 60 times and following outcomes were observed.

Side	1	2	3	4	5	6	Total
Frequency	8	9	13	7	15	8	n = 60

Test the hypothesis that the die is fair i.e., all sides have equal chance of appearing against the die is unfair at 5 % level of significance.

Solution:

Data

X = Face value of die (1, 2, 3, 4, 5, 6)

No. of categories (k) = 6

Sample size (n) = 60

Null and Alternative Hypothesis

H_0: Die is fair i.e frequencies are according to uniform distribution

H_1: Die is not fair i.e., frequencies are not according to uniform distribution

Level of significance

Given level of significance = 0.05

Test Statistics

The difference between the theoretical cumulative distribution function F_T (x) and the observed or empirical cumulative distribution function F_O (x) is measured by D statistic.

For the discrete distribution, the test statistic is given by,

$$D_i = Max\left|F_O(x_i) - F_T(x_i)\right| \text{ for each i}$$

Calculation of D statistics

Categories (Side)	f_0	fe	Observed cumm. frequency	Theoretical cumm. frequency	Observed relative cumm. frequency $F_O(x_i)$	Theoretical relative cumm. frequency $F_T(x_i)$	D_i
1	8	10	8	10	0.1333	0.1667	0.0334
2	9	10	17	20	0.2833	0.3333	0.05
3	13	10	30	30	0.50	0.50	0
4	7	10	37	40	0.6167	0.6667	0.05
5	15	10	52	50	0.8667	0.8333	0.0334
6	8	10	60	60	1	1	
Total	60	60					

Thus, Cal D = 0.05

Tabulated D

level of significance = 0.05

Sample size (n) = 60

General Formula

The critical value for a K-S test at α = 0.05 is:

$$D_{critical} = \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{60}} = \frac{1.36}{7.746} = 0.176$$

Statistical Decision:

Since cal D (0.05) < critical D (0.176), we do not reject null hypothesis H_0 at 5 % level of significance.

Conclusion:

The die is fair i.e., observed frequencies are according to uniform distribution.

Critical Values Table for One-Sample K-S Test

The critical values for the Kolmogorov-Smirnov (K-S) test depend on the sample size (n) and the significance level (α). Below is a simplified table for commonly used significance levels (α = 0.10, 0.05, 0.01):

Critical Values Table for One-Sample K-S Test is given by,

$$D_{critical} = \frac{C\ (\alpha)}{\sqrt{n}}$$

Where c(α) is a constant based on the significance level.

Significance Level (α)	**C(α)**
0.10	1.22
0.05	1.36
0.01	1.63