Two sample Kolmogorov Smirnov test

K-S test can also be used in two sample situations and the objective is to compare two distributions, and determines whether two samples come from the same distribution or not. Comparing distributions can be either both continuous or both discrete.

Test Assumptions

1. Two samples are drawn randomly from populations.
2. Two comparing distributions can be discrete or continuous

Hypothesis

H_0: Both samples come from a population with the same distribution i.e. $PDF_1 = PDF_2$

H_1: Two samples come from different distributions i.e. $PDF_1 \neq PDF_2$

Test Statistics

Suppose that the first sample has size *m* observations with an observed cumulative distribution function of *F*(*x*) and that the second sample has size *n* observations with an observed cumulative distribution function of *F*(*y*). Test statistic is the maximum absolute difference between two cumulative probability distributions.
So, the test statistic is given by,

$$D = Max\left|F(x) - F(y)\right|$$

Decision Rule: Reject H0 if cal D > Critical D

Large Sample Case

If m and n are sufficiently large the critical value of D i.e. $D_{\alpha,m,n}$ is approximated by following formula,

$$D_{\alpha,m,n} = C(\alpha)\sqrt{\frac{m+n}{m.n}}$$

where *c*(*α*) = the inverse of the Kolmogorov distribution at *α*.

α	*0.10*	*0.05*	*0.025*	*0.01*	*0.005*	*0.001*
C(α)	*1.22*	*1.36*	*1.48*	*1.63*	*1.73*	*1.95*

Example:

Forty boys and fifty girls of similar age group are randomly selected and given an IQ test. The scores thus obtained for boys as well as girls are tabulated below in terms of their frequencies.

Scores	60-65	65-70	70-75	75-80	80-85	85-90	90-95	Total
Frequency Boys	0	2	4	8	15	7	4	40
Frequency Girls	4	6	10	12	8	6	4	50

Perform KS test to known whether there is significant difference between the score for boys and girls i.e., two distributions are same against not.

Solution:

Data

First sample size (m) = 40

Second sample size (n) = 50

There are seven categories.

Hypothesis

H_0 : The distribution of IQ scores of boys and girls are same

H_1 : The distribution of IQ scores of boys and girls are same

Calculated D

IQ Score	Frequency (B) f(x)	Frequency (G) f(y)	Cumm. Frequency (B)	Cumm. Frequency (G)	Cumm. Relative Frequency (B)	Cumm. Relative Frequency (G)	|F(x)-F(y)|
60-65	0	4	0	4	0	0.08	0.08
65-70	2	6	2	10	0.05	0.20	0.15
70-75	4	10	6	20	0.15	0.40	0.25
75-80	8	12	14	32	0.35	0.64	**0.29**
80-85	15	8	29	40	0.725	0.80	0.075
85-90	7	6	36	46	0.90	0.92	0.02
90-95	4	4	40	50	1	1	0

The largest difference between any two cummulative relative frequencies is 0.25, hence cal D = 0.29

Critical D

The critical D is approximated by the following formula.

$$D_{\alpha,m,n} = C(\alpha)\sqrt{\frac{m+n}{m.n}}$$

In our case, m = 40, n = 50 and α =0.05, the critical D is given by,

$$D_{0.05,40,50} = 1.36\sqrt{\frac{40+50}{40*50}} = 0.288$$

Statistical Decision

Since, cal D = 0.29 > Critical D = 0.288, the null hypothesis is rejected.