Kolmogorov-Smirnov (K-S) test

1. ตั้งสมมติฐาน

H_0 : ทั้ง 2 sample มาจาก **distribution เดียวกัน**

H_1 : ทั้ง 2 sample มาจาก**คนละ distribution**

2. คำนวณ empirical distribution function จากทั้ง 2 sample

$$F_X(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(x_i \le x) \quad ; \quad m = \text{จำนวนข้อมูลใน sample ที่ } 1$$

$$F_Y(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(y_j \le x) \quad ; \quad n = \text{จำนวนข้อมูลใน sample ที่ } 2$$

3. คำนวณระยะห่างที่มากที่สุดของทั้ง 2 sample ของแต่ละ sample point

$$D = \sup_x |F_X(x) - F_Y(x)|$$

4. คำนวณค่าระยะห่างที่จุด critical point ของ K-S test (α = significant level)

$$D_c = c(\alpha)\sqrt{\frac{nm}{n+m}} \quad ; \quad c(\alpha) = \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}$$

5. สรุปผลโดยพิจารณาว่า

- ถ้าระยะห่างที่มากที่สุดจากข้อ 3 $>$ ค่าที่คำนวณได้ทางสถิติจากข้อ 4: ปฏิเสธ H0 หมายความว่าทั้ง **2 sample มาจากคนละ distribution**
- ถ้าระยะห่างที่มากที่สุดจากข้อ 3 $\le$ ค่าที่คำนวณได้ทางสถิติจากข้อ 4: ไม่ปฏิเสธ H0 หมายความว่าทั้ง **2 sample มาจาก distribution เดียวกัน**

<u>ตัวอย่างการคำนวณ</u>

Sample 1: $X = [2.1, 2.3, 2.5, 2.7, 2.9]$

Sample 2: $Y = [1.8, 2.0, 2.4, 2.6, 3.0]$

1. ตั้งสมมติฐาน

   H_0 : ทั้ง 2 sample มาจาก **distribution เดียวกัน**

   H_1 : ทั้ง 2 sample มาจาก**คนละ distribution**

2. คำนวณ empirical distribution function จากทั้ง 2 sample

   สำหรับ x=1.8

$$x_1 = 2.1, x_2 = 2.3, x_3 = 2.5, x_4 = 2.7, x_5 = 2.9$$

$$F_X(1.8) = \frac{1}{5}\begin{pmatrix} \mathbf{1}(x_1 \leq 1.8) + \mathbf{1}(x_2 \leq 1.8) \\ +\mathbf{1}(x_3 \leq 1.8) + \mathbf{1}(x_4 \leq 1.8) + \mathbf{1}(x_5 \leq 1.8) \end{pmatrix}$$

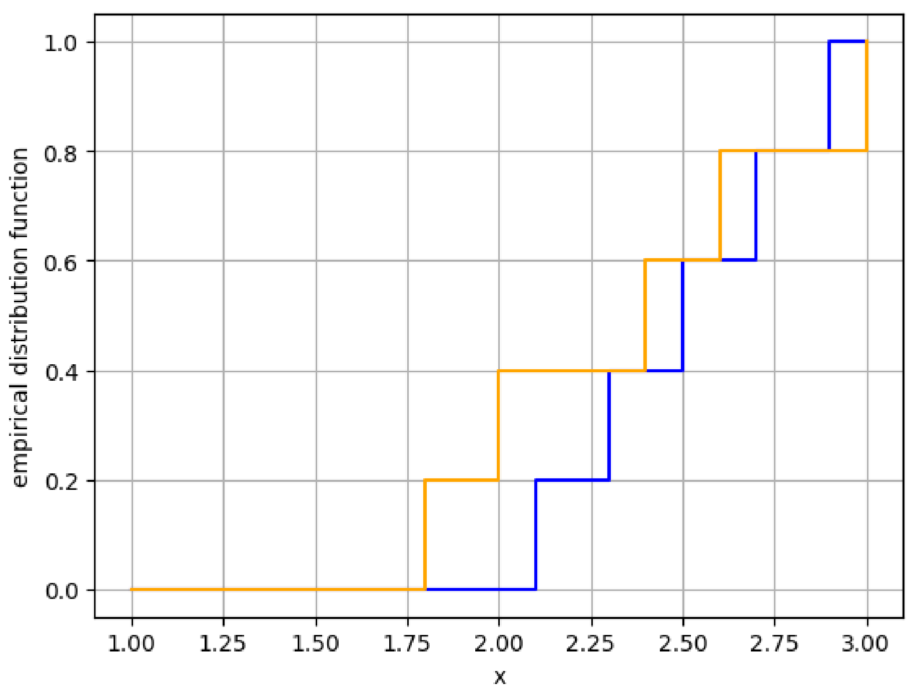
$$= \frac{1}{5}(0 + 0 + 0 + 0 + 0)$$
$$= 0$$

$$y_1 = 1.8, y_2 = 2.0, y_3 = 2.4, y_4 = 2.6, y_5 = 3.0$$

$$F_Y(1.8) = \frac{1}{5}\begin{pmatrix} \mathbf{1}(y_1 \leq 1.8) + \mathbf{1}(y_2 \leq 1.8) + \mathbf{1}(y_3 \leq 1.8) \\ +\mathbf{1}(y_4 \leq 1.8) + \mathbf{1}(y_5 \leq 1.8) \end{pmatrix}$$

$$= \frac{1}{5}(1 + 0 + 0 + 0 + 0)$$
$$= \frac{1}{5} = 0.2$$

x	$F_X(x)$	$F_Y(x)$
1.8	0	0.2
2.0	0	0.4
2.1	0.2	0.4
2.3	0.4	0.4
2.4	0.4	0.6
2.5	0.6	0.6
2.6	0.6	0.8
2.7	0.8	0.8
2.9	1	0.8
3.0	1	1



3. คำนวณระยะห่างที่มากที่สุดของทั้ง 2 sample ของแต่ละ sample point

x	$F_X(x)$	$F_Y(x)$	$\lvert F_X(x) - F_Y(x)\rvert$
1.8	0	0.2	$\left\lvert 0 - \frac{1}{5}\right\rvert = \frac{1}{5} = 0.2$
2.0	0	0.4	0.40
2.1	0.2	0.4	0.20
2.3	0.4	0.4	0.00
2.4	0.4	0.6	0.20
2.5	0.6	0.6	0.00
2.6	0.6	0.8	0.20
2.7	0.8	0.8	0.00
2.9	1	0.8	0.20
3.0	1	1	0.00

$$D = \sup_x \lvert F_X(x) - F_Y(x)\rvert = 0.4$$

4. คำนวณค่าระยะห่างที่จุด critical point ของ K-S test

   กำหนดให้ $\alpha = 0.05$

$$\begin{aligned} D_c &= c(\alpha)\sqrt{\frac{nm}{n+m}} \quad ; \quad c(\alpha) = \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)} \\ &= \sqrt{-\frac{1}{2}\ln\left(\frac{0.05}{2}\right)}\sqrt{\frac{5\cdot 5}{5+5}} \\ &= 0.86 \end{aligned}$$

5. สรุปผลได้ว่า

$$D = 0.4 < 0.86 = D_C$$

   ซึ่งหมายความว่าเราจะไม่ปฏิเสธ H0 หมายความว่าทั้ง **2 sample มาจาก distribution เดียวกัน**