

QUY TRÌNH KIỂM ĐỊNH K-S, KOLMOGOROV-SMIRNOV HAI MẪU

Ý nghĩa của kiểm định K-S, Kolmogorov-Smirnov hai mẫu:

+ Kiểm định Kolmogorov - Smirnov hai mẫu được sử dụng để kiểm tra xem hai mẫu có đến từ cùng một phân phối hay không?

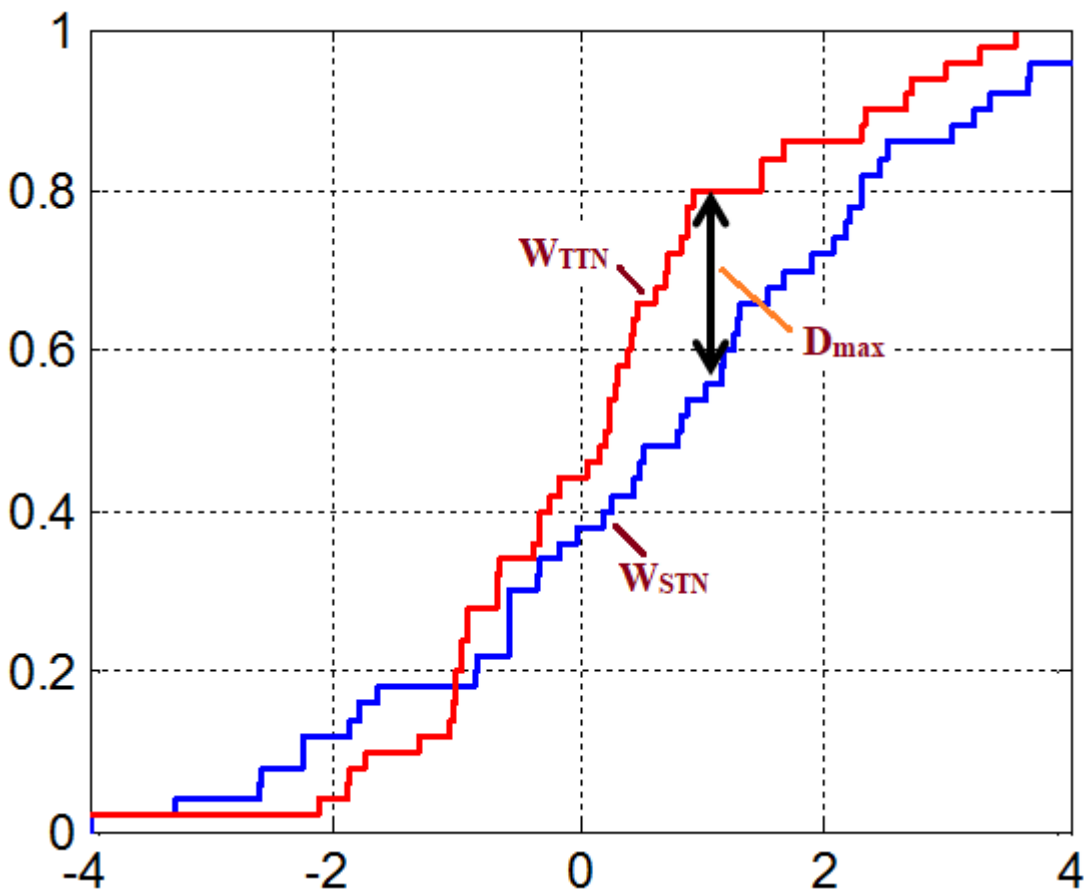
+ Kiểm định Kolmogorov - Smirnov hai mẫu được sử dụng trong so sánh đánh giá tính tối ưu của phân phối A với phân phối B.

Giả sử rằng mẫu đầu tiên có kích thước m với hàm phân phối tích lũy quan sát là $F(x)$ và mẫu thứ hai có kích thước n với hàm phân phối tích lũy quan sát được của $G(x)$.

Định nghĩa:

$$F(x) = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2\pi^2/(8x^2)}$$

Giả sử đồ thị của hàm W_{TTN} (Màu đỏ) và hàm W_{STN} (Màu xanh)



Gọi: $\{D\}_{m,n} = \{D\}_{\text{MAX}} = \underset{x}{\mathop{\max}} \left| \{F\}_m(x) - \{G\}_n(x) \right|$ Cho m và n đủ lớn: $\{D\}_{m,n,\alpha}$ là giá trị tới hạn.

Giả thuyết không là H_0 : Cả hai mẫu đều đến từ một quần thể có cùng phân phối. Đối với phép thử Kolmogorov - Smirnov hai mẫu về tính quy tắc, Chúng tôi bác bỏ giả thuyết không (ở mức ý nghĩa α) nếu $D_{m,n} = D_{max} > D_{m,n,\alpha}$ trong đó $D_{m,n,\alpha}$ là giá trị tới hạn.

1. Tìm $D_{\max}$

Lập bảng Tần suất tích lũy (Phân phối tích lũy) như bảng sau trong EXCEL.

	A	B	C	D	E	F	G	H	I
1	Bảng tần xuất tích lũy của hai mẫu TTN và STN								
2		TTN	STN	Cum % TTN	Cum % STN	diff	alpha	0.05	
3	Không khả thi	2	5	0,0103	0,0253	0,0150			
4	ít khả thi	5	10	0,0359	0,0758	0,0399			
5	TB	36	16	0,2205	0,1566	0,0639			
6	Khá khả thi	115	68	0,8103	0,5000	0,3103			
7	Rất khả thi	37	99	1	1	0			
8	Tổng	195	198						
9					D-MAX	0,31025641			
10					D-m,n, α	0,1360			
11									
12									

Trong đó:

Cột D chứa phân phối tích lũy cho TTN (dựa trên cột B), cột E chứa phân phối tích lũy cho STN và cột F chứa giá trị tuyệt đối của hiệu các cột D và E.

Cụ thể:

- Cột D, ô D3 chứa công thức = B3 / B8, ô D4 chứa công thức = B4 / B8 + D3; ô D5 chứa công thức = B5/B8+D4;...
- Tương tự cột D cho cột E.
- Cột F, ô F3 chứa công thức = ABS (D3-E3); F4 chứa công thức=ABS(D4-F4);...
- Ô F9 chứa công thức = max (F3:F7).
- Ô F10 là giá trị $D_{m,n,\alpha}$ được tính ở mục 2 sau.

Vậy: $D_{\max}$ =giá trị ô F9.

2. Tính $\{D_{m,n,\alpha}\}$

a) Trường hợp 1: m,n có trong bảng Kolmogorov - Smirnov hai mẫu;

$\{D_{m,n,\alpha}\} = \frac{c(\alpha)}{m.n}$; Với $c(\alpha)$ được tra bảng với giá trị α , dòng m, cột n tương ứng.

Ví dụ: m=10, n=15, $\alpha=0,05$

$\alpha = .05$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	-	-	-	-	-	-	16	18	20	22	24	26	26	28	30	32	34	36	38
3	-	-	-	15	18	21	21	24	27	30	30	33	36	36	39	42	45	45	48
4	-	-	16	20	20	24	28	28	30	33	36	39	42	44	48	48	50	53	60
5	-	15	20	25	24	28	30	35	40	39	43	45	46	55	54	55	60	61	65
6	-	18	20	24	30	30	34	39	40	43	48	52	54	57	60	62	72	70	72
7	-	21	24	28	30	42	40	42	46	48	53	56	63	62	64	68	70	76	79
8	16	21	28	30	34	40	48	46	48	53	60	62	64	67	80	77	80	82	88
9	18	24	28	35	39	42	46	54	53	59	63	65	70	75	78	82	92	89	93
10	20	27	30	40	40	46	48	53	70	60	66	70	74	80	84	89	92	94	110
11	22	30	33	39	43	48	53	59	60	77	72	75	8	84	89	93	97	102	107
12	24	30	36	43	48	53	60	63	66	72	84	81	86	93	96	100	108	108	116
13	26	33	39	45	52	56	62	65	70	75	81	91	89	96	101	105	110	114	120
14	26	36	42	46	54	63	64	70	74	82	86	89	112	98	106	111	116	121	126
15	28	36	44	55	57	62	67	75	80	84	93	96	98	120	114	116	123	127	135

$\Rightarrow c(0.05) = 80; \{D\}_{10,15,0.05} = \frac{80}{\sqrt{10 \cdot 15}} = 0,5333$

Bảng Kolmogorov - Smirnov hai mẫu

$\alpha = .05$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	-	-	-	-	-	-	16	18	20	22	24	26	26	28	30	32	34	36	38
3	-	-	-	15	18	21	21	24	27	30	30	33	36	36	39	42	45	45	48
4	-	-	16	20	20	24	28	28	30	33	36	39	42	44	48	48	50	53	60
5	-	15	20	25	24	28	30	35	40	39	43	45	46	55	54	55	60	61	65
6	-	18	20	24	30	30	34	39	40	43	48	52	54	57	60	62	72	70	72
7	-	21	24	28	30	42	40	42	46	48	53	56	63	62	64	68	70	76	79
8	16	21	28	30	34	40	48	46	48	53	60	62	64	67	80	77	80	82	88
9	18	24	28	35	39	42	46	54	53	59	63	65	70	75	78	82	92	89	93
10	20	27	30	40	40	46	48	53	70	60	66	70	74	80	84	89	92	94	110
11	22	30	33	39	43	48	53	59	60	77	72	75	8	84	89	93	97	102	107
12	24	30	36	43	48	53	60	63	66	72	84	81	86	93	96	100	108	108	116
13	26	33	39	45	52	56	62	65	70	75	81	91	89	96	101	105	110	114	120
14	26	36	42	46	54	63	64	70	74	82	86	89	112	98	106	111	116	121	126
15	28	36	44	55	57	62	67	75	80	84	93	96	98	120	114	116	123	127	135

b) Trường hợp 2: m, n lớn vượt quá bảng Kolmogorov - Smirnov hai mẫu.

Ta có: $\{D\}_{m,n,\alpha} = c(\alpha) \cdot \sqrt{\frac{m+n}{m \cdot n}}$

Trong đó: $c(\alpha)$ là số **cuối cùng** của tử số trong bảng Kolmogorov - Smirnov một mẫu dưới đây.

Bảng Kolmogorov - Smirnov một mẫu

$n \backslash \alpha$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	1.94947	1.62762	1.51743	1.35810	1.22385	1.13795	1.07275
	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$

Ví dụ: Giả sử: $m=195$; $n=198$; $\alpha=0.05 \Rightarrow c(\alpha)=1.35810$.

Với $m=195$, $n=198$ thì $\{D\}_{m,n,\alpha} = \{D\}_{195,198,0.05} = 1.3581 \cdot \sqrt{\frac{195+198}{195 \cdot 198}} = 0.1360$

3. Thực hành kiểm định Kolmogorov - Smirnov hai mẫu

Giả sử ta có hai phân phối tích lũy (như bảng sau).

	A	B	C	D	E	F	G	H	I
1	Bảng tần suất tích lũy của hai mẫu TTN và STN								
2		TTN	STN	Cum % TTN	Cum % STN	diff	alpha	0.05	
3	Không khả thi	2	5	0,0103	0,0253	0,0150			
4	Ít khả thi	5	10	0,0359	0,0758	0,0399			
5	TB	36	16	0,2205	0,1566	0,0639			
6	Khá khả thi	115	68	0,8103	0,5000	0,3103			
7	Rất khả thi	37	99	1	1	0			
8	Tổng	195	198						
9					D-MAX	0,31025641			
10					D-m,n, α	0,1360			
11									
12									

Từ bảng trên ta có $D_{\max}=0,3102$.

Sử dụng kiểm định Kolmogorov-Smirnov hai mẫu, với mức $\alpha = 0,05$.

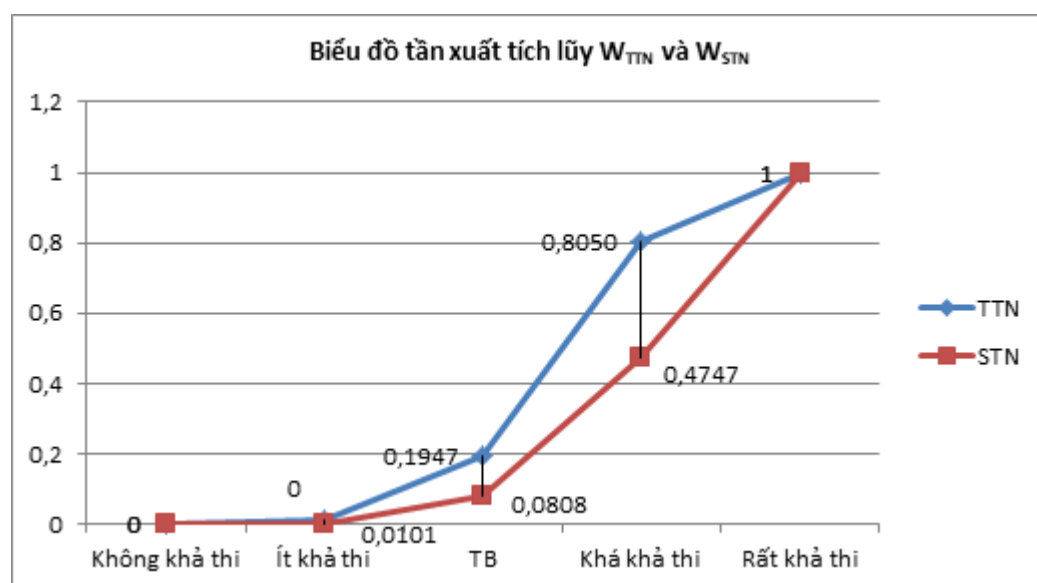
Tra bảng Kolmogorov - Smirnov với: $m = 195; n=198$, cho ta kết quả là:

$$C(\alpha) = 1,3581 \Rightarrow \{D\}_{m,n,\alpha} = c(\alpha) \sqrt{\frac{m+n}{mn}} = 1.3581 \sqrt{\frac{195+198}{195 \cdot 198}} = 0.13605$$

$$\Rightarrow D_{m,n,\alpha} = 0.1360 < D_{\max} = 0,3302; \text{ với mức ý nghĩa sai số } 5\%$$

Do đó, giả thuyết H_0 bị bác bỏ.

Ta có biểu đồ tần suất tích lũy như sau:



Đánh giá: Biểu đồ Phân phối W_{STN} nằm phía bên phải biểu đồ phân phối W_{TTN} . Do vậy phân phối W_{STN} tốt hơn phân phối W_{TTN} với mức ý nghĩa 95%.

Chúc các bạn thành công.

Nguồn: <http://www.real-statistics.com>, <http://www.real-statistics.com>