Kolmogorov-Smirnov test (KS-test)

Teorijske osnove

Normalitet distribucija varijabli, tj. sličnost empirijskih distribucija s normalnom distribucijom je uvjet za korištenje mnogih statističkih metoda.

Veličina odstupanja empirijske distribucije od normalne distribucije može se testirati statističkim postupcima kao što su *Kolmogorov-Smirnov test*, *Lilliefors test* i *Shapiro-Wilk W test*.

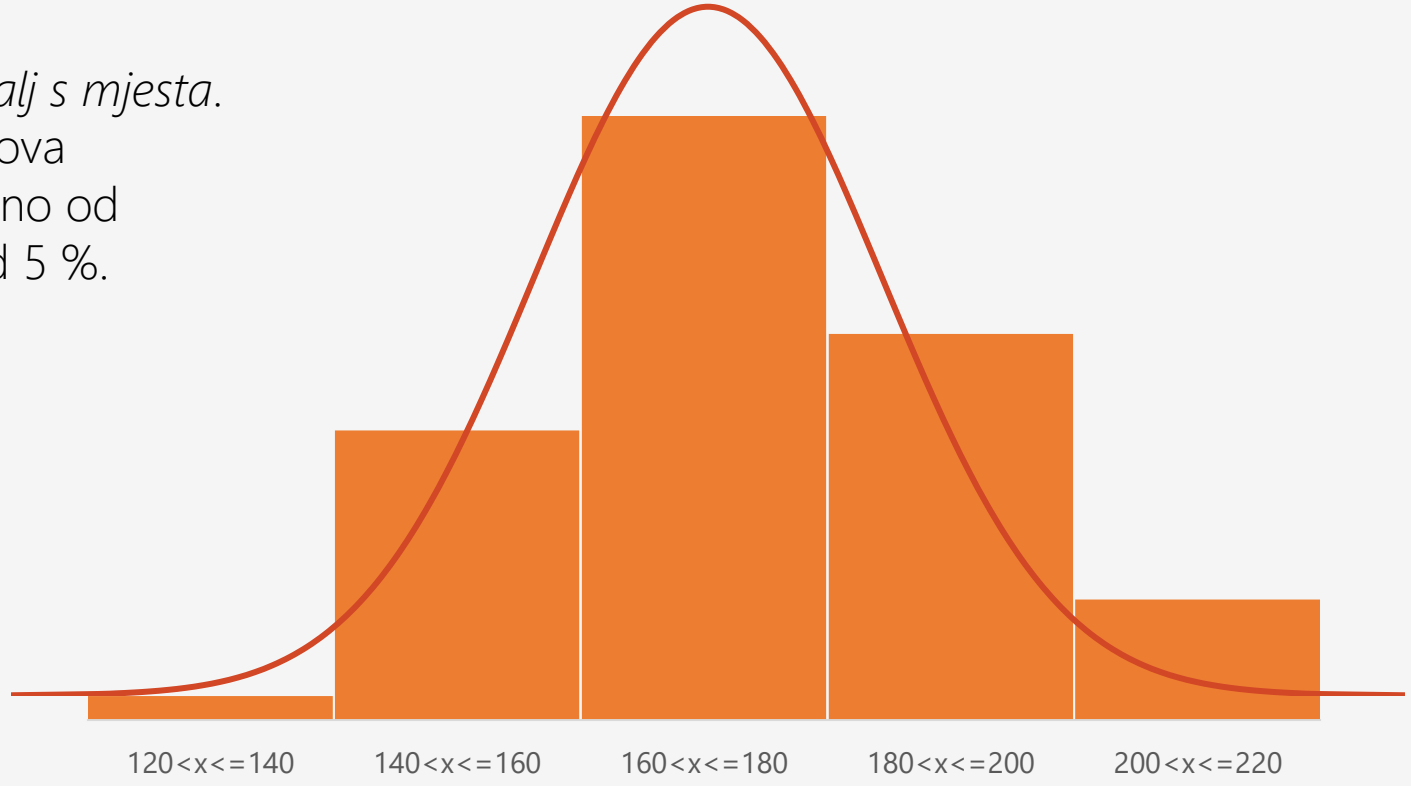
Oblik empirijske distribucije može se opisati *mjerama asimetrije i izduženosti distribucije*.

Teorijske osnove

Kolmogorov-Smirnovljev test - postupak za utvrđivanje normaliteta neke empirijske distribucije.

Temelji se na usporedbi *empirijskih relativnih kumulativnih frekvencija* (*rcf*) i *teoretskih relativnih kumulativnih frekvencija* (*trcf*).

Primjer: 60 entiteta izmjereno je testom *skok udalj s mjesta*. Potrebno je uz pomoć *KS-testa* utvrditi da li njihova (empirijska) distribucija odstupa statistički značajno od (teoretske) normalne distribucije uz pogrešku od 5 %.



Teorijske osnove

1. Odrediti prikladan broj razreda i njihovu veličinu (interval razreda) te u njih grupirati entitete.

Intervali razreda	f
$120 < x \leq 140$	1
$140 < x \leq 160$	12
$160 < x \leq 180$	26
$180 < x \leq 200$	16
$200 < x \leq 220$	5

Teorijske osnove

2. Na temelju grupiranih podataka izračunati empirijske kumulativne i relativne kumulativne frekvencije.

Intervali razreda	f	cf	rcf
$120 < x \leq 140$	1	1	0,0167
$140 < x \leq 160$	12	13	0,2167
$160 < x \leq 180$	26	39	0,6500
$180 < x \leq 200$	16	55	0,9167
$200 < x \leq 220$	5	60	1

Teorijske osnove

3. Na temelju aritmetičke sredine (177,25) i standardne devijacije (16,86), standardizirati gornje granice svakog razreda.

Intervali razreda	f	cf	rcf	z
$120 < x \leq 140$	1	1	0,0167	-2,21
$140 < x \leq 160$	12	13	0,2167	-1,02
$160 < x \leq 180$	26	39	0,6500	0,16
$180 < x \leq 200$	16	55	0,9167	1,35
$200 < x \leq 220$	5	60	1	2,54

Teorijske osnove

4. Uz pomoć tablice površine ispod normalne distribucije izračunati površinu od lijevog kraja krivulje do određene *z-vrijednosti*, odnosno izračunati teoretske relativne kumulativne frekvencije.

Intervali razreda	f	cf	rcf	z	trcf
$120 < x \leq 140$	1	1	0,0167	-2,21	0,0135
$140 < x \leq 160$	12	13	0,2167	-1,02	0,1531
$160 < x \leq 180$	26	39	0,6500	0,16	0,5648
$180 < x \leq 200$	16	55	0,9167	1,35	0,9114
$200 < x \leq 220$	5	60	1	2,54	0,9944

Teorijske osnove

5. Izračunati odstupanja između empirijske i teoretske relativne kumulativne frekvencije.

Intervali razreda	f	cf	rcf	z	trcf	D
$120 < x \leq 140$	1	1	0,0167	-2,21	0,0135	0,0032
$140 < x \leq 160$	12	13	0,2167	-1,02	0,1531	0,0636
$160 < x \leq 180$	26	39	0,6500	0,16	0,5648	0,0852
$180 < x \leq 200$	16	55	0,9167	1,35	0,9114	0,0053
$200 < x \leq 220$	5	60	1	2,54	0,9944	0,0056

Teorijske osnove

6. Odrediti najveće odstupanje empiriske i teoretske relativne kumulativne (*maxD*) frekvencije i usporediti ga sa tabličnom vrijednošću *KS-testa* određenom za odgovarajući broj entiteta.

Intervali razreda	f	cf	rcf	z	trcf	D
$120 < x \leq 140$	1	1	0,0167	-2,21	0,0135	0,0032
$140 < x \leq 160$	12	13	0,2167	-1,02	0,1531	0,0636
$160 < x \leq 180$	26	39	0,6500	0,16	0,5648	**0,0852**
$180 < x \leq 200$	16	55	0,9167	1,35	0,9114	0,0053
$200 < x \leq 220$	5	60	1	2,54	0,9944	0,0056

Teorijske osnove

n	p=0,05
1	0,975
2	0,842
3	0,708
4	0,624
5	0,563
6	0,519
7	0,483
8	0,454
9	0,430
10	0,409
11	0,391
12	0,375
13	0,361

n	p=0,05
14	0,349
15	0,338
16	0,327
17	0,318
18	0,309
19	0,301
20	0,294
21	0,287
22	0,281
23	0,275
24	0,269
25	0,264
26	0,259

n	p=0,05
27	0,254
28	0,250
29	0,246
30	0,242
35	0,224
40	0,210
45	0,198
50	0,188
55	0,180
60	**0,172**
65	0,166
70	0,160
75	0,154

n	p=0,05
80	0,150
85	0,145
90	0,141
95	0,137
100	0,134

$$n > 100$$

$$KS - test_{0,05} = \frac{1,36}{\sqrt{n}}$$

Teorijske osnove

maxD < KS-test	→	H_0 - distribucija ne odstupa statistički značajno od *normalne*
maxD > KS-test	→	H_1- distribucija statistički značajno odstupa od *normalne*
0,0852 < 0,172	→	H_0- distribucija ne odstupa statistički značajno od *normalne*

Microsoft Excel

Zadatak 1: Utvrdite da li se empirijska distribucija varijable SDM koja se nalazi u datoteci JUDO.xls statistički značajno razlikuje od normalne ili Gaussova distribucije uz pogrešku 0,05.