Kolmogorov-Smirnov Testi

Bu bölümde tek örneklem ve iki örneklem K-S testi anlatılacaktır.

Tek Örneklem KS Testi

$X_1, X_2, \dots, X_n$ rastgele örnekleminin belirlenen bir dağılımdan (Düzgün, Üstel, Normal, v.b.) gelip gelmediğini test etmek amacıyla kullanılan bir uyum iyiliği testidir.

Sıfır ve alternatif hipotezleri

$$H_O: F(x) = F_0(x)$$

$$H_1: F(x) \neq F_0(x)$$

olarak ifade edilir. Burada, F_0 belirlenen dağılım fonksiyonunu gösterir. F ise rastgele örneklemin geldiği dağılımın dağılım fonksiyonunu temsil eder.

Test İstatistiği:

$$D_n = max\{max|F_0(x_i) - S_n(x_i)|, max|F_0(x_i) - S_n(x_{i-1})|\}$$

olarak verilmiştir.

Burada,

$$S_n(x) = \frac{\#(i: X_i < x)}{n}$$

şeklinde tanımlanır.

$\#(\cdot)$: Parantez içindeki şartı sağlayan durumların sayısını gösterir.

Karar: Tablo değeri $k_{\alpha,n}$ olmak üzere; Eğer $D_n > k_{\alpha,n}$ ise H_0 hipotezi reddedilir.

Örnek: Aşağıdaki veri setinin dağılımının $N(0,1)$ olup olmadığını Kolmogorov-Smirnov testini kullanarak sınayınız.

$$X_i: -0.476, \quad 0.184, \quad -1.362, \quad 0.862, \quad 0.455$$

Sıfır Hipotezi

$$H_0: Veri\ setinin\ dağılımı\ N(0,1)'dir.$$

veya

$$H_0: F(x) = \Phi(x)$$

olarak ifade edilir. Burada, $\Phi(x)$ $N(0,1)$ dağılımının dağılım fonksiyonunu gösterir.

Çözüm:

1. Gözlemler küçükten büyüğe doğru sıralanır.

-1.362 -0.476 0.184 0.455 0.862

2) $S_n(x)$ değerleri hesaplanır;

$$S_n(x) = \begin{cases} 0, & x < -1.362 \\ 1/5, & -1.362 \le x < -0.476 \\ 2/5, & -0.476 \le x < 0.184 \\ 3/5, & 0.184 \le x < 0.455 \\ 4/5, & 0.455 \le x < 0.862 \\ 1, & 0.862 \le x \end{cases}$$

3) Aşağıdaki tablodan yararlanarak D_n istatistiği hesaplanır.

x_i	$F_0(x_i)$	$S_n(x_i)$	$\lvert F_0(x_i) - S_n(x_i)\rvert$	$\lvert F_0(x_i) - S_n(x_{i-1})\rvert$
-1.362	0.0866	0.2	0.1134	0.0866
-0.476	0.3170	0.4	0.0830	0.1170
0.184	0.5729	0.6	0.0271	**0.1729**
0.455	0.6755	0.8	0.1245	0.0755
0.862	0.8057	1	**0.1943**	0.0057

$$D_n = max\{0.1943, 0.1729\} = 0.1943$$

4) $\alpha = 0.05$ ve $n = 5$ için tablo değeri $k_{\alpha,n} = 0.563$ olup $D_n < k_{\alpha,n}$ olduğundan H_0 hipotezi reddedilemez.

Sonuç: Veri seti $N(0,1)$ dağılımından gelmektedir.

Örnek: Aşağıdaki veri setinin dağılımının $U(0,4)$ olup olmadığını Kolmogorov-Smirnov testini kullanarak sınayınız.

0.06 0.17 0.24 0.38 0.52 0.67 0.93 0.96 1.41 1.55

1.61 1.80 2.30 2.59 2.60 2.92 3.12 3.28 3.76 3.82

$$H_0: Veri\ setinin\ dağılımı\ U(0,4)'dır.$$

$$H_1: Veri\ setinin\ dağılımı\ U(0,4)\ değildir.$$

Çözüm:

$U(0,4)$ dağılımının dağılım fonksiyonu aşağıdaki gibi bulunur.

$$F_0(x) = \int_0^x f(x)dx = \int_0^x \frac{1}{4}dx = \frac{x}{4}\ , \qquad 0 < x < 4$$

Gözlemler küçükten büyüğe doğru sıralandıktan sonra aşağıdaki tablodan yararlanarak D_n istatistiği hesaplanır.

x_i	$F_0(x_i)$	$S_n(x_i)$	$\lvert F_0(x_i) - S_n(x_i)\rvert$	$\lvert F_0(x_i) - S_n(x_{i-1})\rvert$
0.06	0.0150	0.05	0.0350	0.0150
0.17	0.0425	0.10	0.0575	0.0075
0.24	0.0600	0.15	0.0900	0.0400
0.38	0.0950	0.20	0.1050	0.0550
0.52	0.1300	0.25	0.1200	0.0700
0.67	0.1675	0.30	0.1325	0.0825
0.93	0.2325	0.35	0.1175	0.0675
0.96	0.2400	0.40	**0.1600**	**0.1100**
1.41	0.3525	0.45	0.0975	0.0475
1.55	0.3875	0.50	0.1125	0.0625
1.61	0.4025	0.55	0.1475	0.0975
1.80	0.4500	0.60	0.1500	0.1000
2.30	0.5750	0.65	0.0750	0.0250
2.59	0.6475	0.70	0.0525	0.0025
2.60	0.6500	0.75	0.1000	0.0500
2.92	0.7300	0.80	0.0700	0.0200
3.12	0.7800	0.85	0.0700	0.0200
3.28	0.8200	0.90	0.0800	0.0300
3.76	0.9400	0.95	0.0100	0.0400
3.82	0.9550	1	0.0450	0.0050

$$D_n = \max\{0.16, 0.11\} = 0.16$$

$\alpha = 0.05$ ve $n = 20$ için tablo değeri $k_{\alpha,n} = 0.294$ olup $D_n < k_{\alpha,n}$ olduğundan H_0 hipotezi reddedilemez.

Sonuç: Veri seti U(0,4) dağılımından gelmektedir.

Örnek: Aşağıdaki veri setinin Üstel($\lambda = 2$) dağılımına sahip olup olmadığını Kolmogorov-Smirnov testini kullanarak sınayınız.

0.92 1.84 3.13 3.90 0.41 4.65 3.88 2.43 2.18 2.23 1.53 2.54 2.55 4.09 3.97

$$H_0: Veri\ setinin\ dağılımı\ Üstel(2)\ dir$$

$$H_1: Veri\ setinin\ dağılımı\ Üstel(2)\ değildir.$$

Çözüm: $Üstel(2)$ dağılımının dağılım fonksiyonu aşağıdaki gibi bulunur.

$$F_0(x) = 1 - e^{-\frac{x}{2}},\ \ x > 0$$

Gözlemler küçükten büyüğe doğru sıralandıktan sonra aşağıdaki tablodan yararlanarak D_n istatistiği hesaplanır.

x_i	$F_0(x_i)$	$S_n(x_i)$	$\lvert F_0(x_i) - S_n(x_i)\rvert$	$\lvert F_0(x_i) - S_n(x_{i-1})\rvert$
0.41	0.1836	0.0667	0.1169	0.1836
0.92	0.3679	0.1333	0.2346	0.3013
1.53	0.5351	0.2000	0.3351	0.4017
1.84	0.6020	0.2667	**0.3353**	**0.4020**
2.18	0.6637	0.3333	0.3303	0.3970
2.23	0.6727	0.4000	0.2727	0.3394
2.43	0.7039	0.4667	0.2372	0.3039
2.54	0.7195	0.5333	0.1862	0.2529
2.55	0.7211	0.6000	0.1211	0.1878
3.13	0.7907	0.6667	0.1240	0.1907
3.88	0.8562	0.7333	0.1229	0.1895
3.90	0.8578	0.8000	0.0578	0.1245
3.97	0.8629	0.8667	0.0038	0.0629
4.09	0.8705	0.9333	0.0628	0.0038
4.65	0.9021	1	0.0979	0.0313

$$D_n = \max\{0.3353, 0.4020\} = 0.4020$$

$\alpha = 0.05$ ve $n = 15$ için tablo değeri $k_{\alpha,n} = 0.338$ olup $D_n > k_{\alpha,n}$ olduğundan H_0 hipotezi reddedilir.

Sonuç: Veri seti Üstel(2) dağılımından gelmemektedir.

İki Örneklem KS Testi

$X_1, X_2, \dots, X_n$ rastgele örneklemi ile $Y_1, Y_2, \dots, Y_n$ rastgele örnekleminin aynı dağılımdan gelip gelmediğini test etmek amacıyla kullanılır.

Sıfır ve alternatif hipotezleri

$$H_O: F_X(z) = F_Y(z)$$

$$H_1: F_X(z) \neq F_Y(z)$$

olarak ifade edilir. Burada, F_X ve F_Y sırasıyla birinci ve ikinci örneklemin dağılım fonksiyonunu gösterir.

Test İstatistiği:

$$D_n = max\{|S_X(z_i) - S_Y(z_i)|\}$$

olarak verilmiştir. Burada,

$$S_X(z) = \frac{\#(i: X_i < z)}{n}$$

$$S_Y(z) = \frac{\#(i: Y_i < z)}{n}$$

ve

$\#(\cdot)$: Parantez içindeki şartı sağlayan durumların sayısı

olarak tanımlanır.

Karar: Tablo değeri $k_{\alpha,n}$ olmak üzere; Eğer $D_n > k_{\alpha,n}$ ise H_0 hipotezi reddedilir.

Örnek: Aşağıdaki veri setlerinin aynı dağılıma sahip olup olmadıklarını Kolmogorov-Smirnov testini kullanarak sınayınız.

$$X_i: 12.3, \quad 16.4, \quad 17.8, \quad 18.6, \quad 20.9$$

$$Y_i: 10.1, \quad 11.5, \quad 13.2, \quad 14.6, \quad 15.3$$

Sıfır ve alternatif hipotezleri

$$H_0: Veri\ setleri\ aynı\ dağılımdan\ gelmektedir.$$

$$H_1: Veri\ setleri\ farklı\ dağılımlardan\ gelmektedir.$$

veya

$$H_O: F_X(z) = F_Y(z)$$

$$H_1: F_X(z) \neq F_Y(z)$$

olarak ifade edilir.

Çözüm: Aşağıdaki tablodan yararlanarak D_n istatistiği hesaplanır.

$X_{(i)}$	$Y_{(i)}$	$S_X(z_i)$	$S_Y(z_i)$	$\lvert S_X(z_i) - S_Y(z_i)\rvert$
-	10.1	0	1/5	0.2
-	11.5	0	2/5	0.4
12.3	-	1/5	2/5	0.2
-	13.2	1/5	3/5	0.4
-	14.6	1/5	4/5	0.6
-	15.3	1/5	1	**0.8**
16.4	-	2/5	1	0.6
17.8	-	3/5	1	0.4
18.6	-	4/5	1	0.2
20.9	-	1	1	0

$D_n = 0.8$ olarak hesaplanmıştır.

$\alpha = 0.05$ ve $n = 5$ için tablo değeri $k_{\alpha,n} = 0.563$ olup

$D_n > k_{\alpha,n}$ olduğundan H_0 hipotezi reddedilir.

Sonuç: Veri setleri aynı dağılımdan gelmemektedir.