ORIGIN ≡ 1

Kolmogorov-Smirnov Test for Goodness of Fit in an ordered sequence

The Kolmogorov-Smirnov Test is designed to test whether observed counts, frequencies, or continuous numerical values collected along an ordered sequence (Y) conform to that expected for a known probability distribution such as equal probability or Normal distribution, or the like. The test proceeds by considering each value in an ordered sequence Y to represent bins (as in a histogram) for which Observed and Expected Cumulative probabilites ϕ(Y) are calculated. (I use Y instead of X here only to conform to Table 17.4 in Sokal & Rohlf 1995).

Model:

Let Expected Cumulative Probabilities ϕ(Y) be specified according to some model over ordered values in Y.

Assumptions:

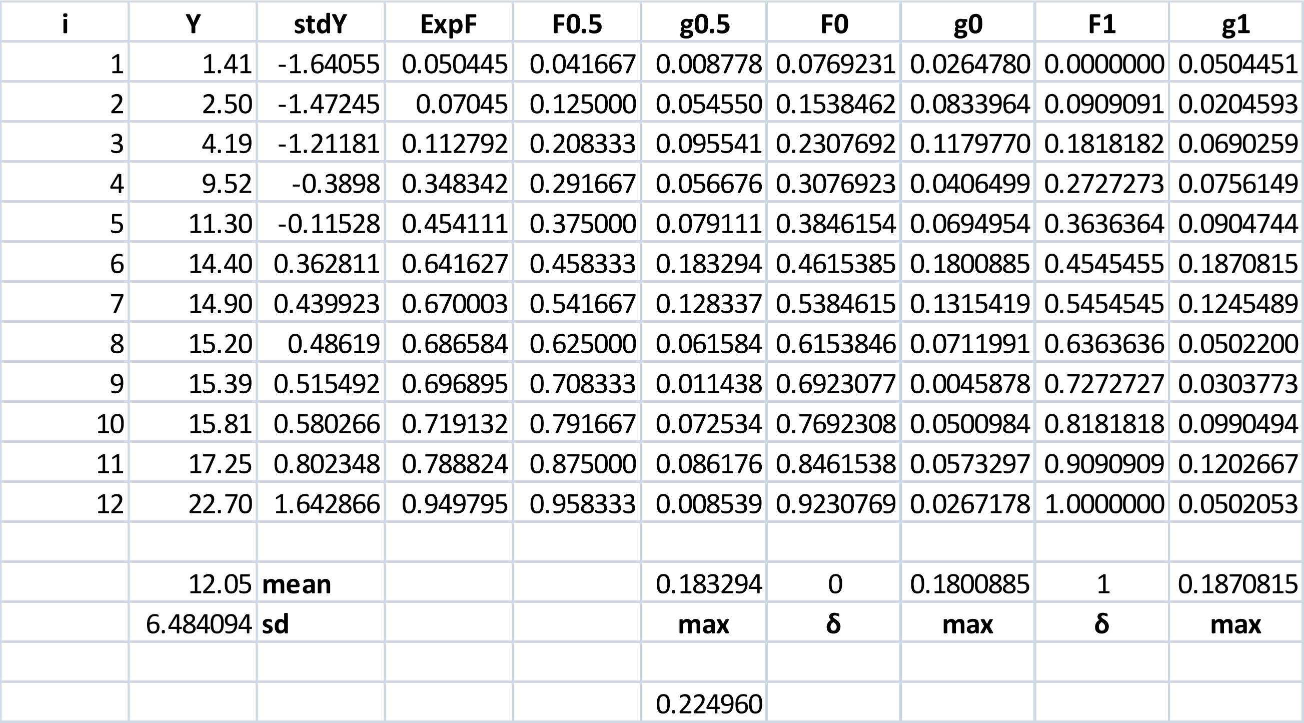
Frequency or numerical data collected for each value of Y are independent of the others.

Hypotheses:

H_0: P_j are distributed according to the model
H_1: P_j differ from the model

< Two sided test

Sokal & Rohlf Table 17.4
based on data in Box 15.2



i	Y	stdY	ExpF	F0.5	g0.5	F0	g0	F1	g1
1	1.41	-1.64055	0.050445	0.041667	0.008778	0.0769231	0.0264780	0.0000000	0.0504451
2	2.50	-1.47245	0.07045	0.125000	0.054550	0.1538462	0.0833964	0.0909091	0.0204593
3	4.19	-1.21181	0.112792	0.208333	0.095541	0.2307692	0.1179770	0.1818182	0.0690259
4	9.52	-0.3898	0.348342	0.291667	0.056676	0.3076923	0.0406499	0.2727273	0.0756149
5	11.30	-0.11528	0.454111	0.375000	0.079111	0.3846154	0.0694954	0.3636364	0.0904744
6	14.40	0.362811	0.641627	0.458333	0.183294	0.4615385	0.1800885	0.4545455	0.1870815
7	14.90	0.439923	0.670003	0.541667	0.128337	0.5384615	0.1315419	0.5454545	0.1245489
8	15.20	0.48619	0.686584	0.625000	0.061584	0.6153846	0.0711991	0.6363636	0.0502200
9	15.39	0.515492	0.696895	0.708333	0.011438	0.6923077	0.0045878	0.7272727	0.0303773
10	15.81	0.580266	0.719132	0.791667	0.072534	0.7692308	0.0500984	0.8181818	0.0990494
11	17.25	0.802348	0.788824	0.875000	0.086176	0.8461538	0.0573297	0.9090909	0.1202667
12	22.70	1.642866	0.949795	0.958333	0.008539	0.9230769	0.0267178	1.0000000	0.0502053
	12.05	mean			0.183294	0	0.1800885	1	0.1870815
	6.484094	sd			max	δ	max	δ	max
					0.224960				

Construct Cumulative Frequencies:

n := 12

i := 1 .. n

SR := READPRN("c:/DATA/Biostatistics/SR15.2R.txt")

$Y := \mathrm{sort}\left(SR^{\langle 3 \rangle}\right)$ **< Ordered values which will be tested against some probability distribution.**

$stdY := \dfrac{Y - \mathrm{mean}(Y)}{\sqrt{\mathrm{Var}(Y)}}$ $\mathrm{mean}(Y) = 12.0475$ $\sqrt{\mathrm{Var}(Y)} = 6.4841$ **< standardizing Y**

ExpF := pnorm(stdY , 0 , 1) **< Expected ϕ(Y) - In this instance, we will be testing a Normal distribution ~N(0,1) for the standardized data stdY.**

$$i = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \end{pmatrix} \quad Y = \begin{pmatrix} 1.41 \\ 2.5 \\ 4.19 \\ 9.52 \\ 11.3 \\ 14.4 \\ 14.9 \\ 15.2 \\ 15.39 \\ 15.81 \\ 17.25 \\ 22.7 \end{pmatrix} \quad stdY = \begin{pmatrix} -1.6406 \\ -1.4724 \\ -1.2118 \\ -0.3898 \\ -0.1153 \\ 0.3628 \\ 0.4399 \\ 0.4862 \\ 0.5155 \\ 0.5803 \\ 0.8023 \\ 1.6429 \end{pmatrix} \quad ExpF = \begin{pmatrix} 0.05045 \\ 0.07045 \\ 0.11279 \\ 0.34834 \\ 0.45411 \\ 0.64163 \\ 0.67 \\ 0.68658 \\ 0.6969 \\ 0.71913 \\ 0.78882 \\ 0.94979 \end{pmatrix}$$

^ standardized Y **^ ϕ(Y) for ~N(0,1)**

Construct Statistics $g_{0.5}$, g_0 & g_1:

$g_{0.5}$ = |ExpF$_i$ - ObsF$_{0.5}$| = absolute value of ExpF minus HKL corrected ObsF$_{0.5}$

g_0 = |ExpF$_i$ - ObsF$_0$| = absolute value of ExpF minus Khamis corrected ObsF$_{0.5}$ where $\delta = 0$

g_1 = |ExpF$_i$ - ObsF$_1$| = absolute value of ExpF minus Khamis corrected ObsF$_{0.5}$ where $\delta = 1$

$$F0.5_i := \frac{(i - 0.5)}{n}$$ **< Harter, Khamis & Lamb correction SR p. 708**

$$g0.5_i := \left| ExpF_i - F0.5_i \right|$$ **< |Exp-Obs$_{0.5}$|**

$$F_{0.5} = \begin{pmatrix} 0.04167 \\ 0.125 \\ 0.20833 \\ 0.29167 \\ 0.375 \\ 0.45833 \\ 0.54167 \\ 0.625 \\ 0.70833 \\ 0.79167 \\ 0.875 \\ 0.95833 \end{pmatrix} \quad g_{0.5} = \begin{pmatrix} 0.0087785 \\ 0.0545502 \\ 0.0955411 \\ 0.0566755 \\ 0.0791108 \\ 0.1832936 \\ 0.1283368 \\ 0.0615837 \\ 0.0114379 \\ 0.0725343 \\ 0.0861758 \\ 0.0085386 \end{pmatrix}$$

$$\delta := 0$$ **< Khamis δ SR p. 711**

$$F0_i := \frac{(i - \delta)}{(n - 2 \cdot \delta + 1)}$$ **< δ_0 correction**

$$g0_i := \left| ExpF_i - F0_i \right|$$ **< |Exp-Obs$_0$| for $\delta = 0$**

$$F_0 = \begin{pmatrix} 0.0769231 \\ 0.1538462 \\ 0.2307692 \\ 0.3076923 \\ 0.3846154 \\ 0.4615385 \\ 0.5384615 \\ 0.6153846 \\ 0.6923077 \\ 0.7692308 \\ 0.8461538 \\ 0.9230769 \end{pmatrix} \quad g_0 = \begin{pmatrix} 0.02647796 \\ 0.08339637 \\ 0.11797697 \\ 0.04064991 \\ 0.06949538 \\ 0.18008847 \\ 0.13154191 \\ 0.07119906 \\ 0.00458775 \\ 0.05009836 \\ 0.05732968 \\ 0.02671779 \end{pmatrix}$$

$\delta := 1$ **< Khamis δ SR p. 711**

$F1_i := \frac{(i - \delta)}{(n - 2 \cdot \delta + 1)}$ **< δ_1 correction**

$g1_i := \left|ExpF_i - F1_i\right|$ **< |Exp-Obs$_1$| for δ = 1**

$$F1 = \begin{pmatrix} 0 \\ 0.0909091 \\ 0.1818182 \\ 0.2727273 \\ 0.3636364 \\ 0.4545455 \\ 0.5454545 \\ 0.6363636 \\ 0.7272727 \\ 0.8181818 \\ 0.9090909 \\ 1 \end{pmatrix} \qquad g1 = \begin{pmatrix} 0.0504451 \\ 0.0204593 \\ 0.0690259 \\ 0.0756149 \\ 0.0904744 \\ 0.1870815 \\ 0.1245489 \\ 0.05022 \\ 0.0303773 \\ 0.0990494 \\ 0.1202667 \\ 0.0502053 \end{pmatrix}$$

Test Statistic d_{max}:

$\max(g0.5) = 0.1832936$

$\max(g0) = 0.1800885$

$\max(g1) = 0.1870815$

$d_{max} := \max(g0.5) + \frac{1}{2 \cdot n}$ $d_{max} = 0.2249603$

Sampling Distribution:

If Assumptions hold and H_0 is true, then D_{max} ~$D_{(\alpha,n)}$

^ Kolmogorov-Smirnov distribution for continuous data in specialized tables (e.g., Zar 2010 Appendix B.9). Sokal & Rohlf offer two different tables depending on whether distribution ExpF is derived from internal estimates or external parameters.

Critical Value of the Test:

$\alpha := 0.05$ **<Type I error must be set**

$C := 0.37543$ **< from Zar 2010 Appendix B.9**

$n = 12$

Decision Rule:

IF d_{max} > C THEN REJECT H_0,OTHERWISE ACCEPT H_0
IF P < α THEN REJECT H_0,OTHERWISE ACCEPT H_0

$d_{max} = 0.225$

Probability Value:

The Kolmogorov-Smirnov distribution is not available for calculating this, so until I find one, we'll have to depend on R's function ks.test() to calculate this for us.

Prototype in R:

```
#KOLMOGOROV-SMIRNOV SINGLE SAMPLE TEST
#FOR CONTINUOUS DATA
#SOKAL & ROHLF TABLE 17.4 & BOX 15.2
SR=read.table("c:/DATA/Biostatistics/SR15.2R.txt")
SR
attach(SR)
Y=sort(bodywt)
stdY=(Y-mean(Y))/sqrt(var(Y))
stdY

ks.test(stdY,pnorm)
```

```
> SR
   gillwt bodywt
1     159  14.40
2     179  15.20
3     100  11.30
4      45   2.50
5     384  22.70
6     230  14.90
7     100   1.41
8     320  15.81
9      80   4.19
10    220  15.39
11    320  17.25
12    210   9.52
```

```
        One-sample Kolmogorov-Smirnov test

data:  stdY
D = 0.225, p-value = 0.5075
alternative hypothesis: two-sided
```

^ Note: in this test, I standardized Y (making stdY) in order to compare stdY with the Normal distribution ~N(0,1)

Kolmogorov-Smirnov Test for Goodness of Fit for Two Samples:

A similar Kolmogorov-Smirnov approach may be used to compare the distribution of one sample with another. According to Sokal & Rohlf, this test is not as sensitive as the Mann-Whitney Test for assessing differences in median. However, it also looks for differences in shapes of the the two distributions.

Construct Chart:

Sokal & Rohlf Biometry 3rd Edition 1995 Example Box 13.9
Testing differences in distribution between Sample A & B:

	length	sample	F1	F2	F1/n1	F2/n2	d
1	100	B	0	1	0.0000	0.1000	0.1000
2	104	A	1	1	0.0625	0.1000	0.0375
3	105	B	1	2	0.0625	0.2000	0.1375
4	107	B	1	3	0.0625	0.3000	0.2375
5	107	B	1	4	0.0625	0.4000	0.3375
6	108	B	1	5	0.0625	0.5000	0.4375
7	109	A	2	5	0.1250	0.5000	0.3750
8	111	B	2	6	0.1250	0.6000	0.4750
9	112	A	3	6	0.1875	0.6000	0.4125
10	114	A	4	6	0.2500	0.6000	0.3500
11	116	A	5	6	0.3125	0.6000	0.2875
12	116	B	5	7	0.3125	0.7000	0.3875
13	118	A	6	7	0.3750	0.7000	0.3250
14	118	A	7	7	0.4375	0.7000	0.2625
15	119	A	8	7	0.5000	0.7000	0.2000
16	120	B	8	8	0.5000	0.8000	0.3000
17	121	A	9	8	0.5625	0.8000	0.2375
18	121	B	9	9	0.5625	0.9000	0.3375
19	123	A	10	9	0.6250	0.9000	0.2750
20	123	B	10	10	0.6250	1.0000	0.3750
21	125	A	11	10	0.6875	1.0000	0.3125
22	126	A	12	10	0.7500	1.0000	0.2500
23	126	A	13	10	0.8125	1.0000	0.1875
24	128	A	14	10	0.8750	1.0000	0.1250
25	128	A	15	10	0.9375	1.0000	0.0625
26	128	A	16	10	1.0000	1.0000	0.0000
			16	10			0.4750
			n1	**n2**			**max d**

F1 & F2 are cumulative counts respectively for samples A & B.
n1 & n2 are total counts for each sample.
d is the absolute difference: |F1/n1 - F2/n2|

Test Statistic:

$d_{max} := 0.4750$

Sampling Distribution:

If Assumptions hold and H_0 is true, then D_{max} ~$D_{(\alpha)}$

^ Sokal & Rohlf Table W

$\alpha := 0.05$ **<Type I error must be set** $d_{max} = 0.475$

Probability Value:

The Kolmogorov-Smirnov distribution is not available for calculating this, so until I find one, we'll have to depend on R's function ks.test() to calculate this for us.

Decision Rule:

IF P < α THEN REJECT H_0,OTHERWISE ACCEPT H_0

Prototype in R:

```
#KOLMOGOROV-SMIRNOV TWO SAMPLE TEST
#FOR CONTINUOUS DATA
#SOKAL & ROHLF EXAMPLE 13.7
SR=read.table("c:/DATA/Biostatistics/SREX13.7R.txt")
SR
attach(SR)
X=length[sample=="A"]
Y=length[sample=="B"]
ks.test(X,Y)
```

```
> SR
  length sample
1   104    A
2   109    A
3   112    A
4   114    A
5   116    A
6   118    A
7   118    A
8   119    A
9   121    A
10   123    A
11   125    A
12   126    A
13   126    A
14   128    A
15   128    A
16   128    A
17   100    B
18   105    B
19   107    B
20   107    B
21   108    B
22   111    B
23   116    B
24   120    B
25   121    B
26   123    B
```

```
        Two-sample Kolmogorov-Smirnov test

data:  X and Y
D = 0.475, p-value = 0.1244
alternative hypothesis: two-sided

Warning message:
In ks.test(X, Y) : cannot compute correct p-values with ties
```

^ Here Calcuulation of max(d) match hand calculation above and in Sokal & Rohlf (1995). This test is supposed to cover continuous distributions, so there should be no ties. As a result, R gives a warning message.

Systat Output:

```
Categorical values encountered during processing are:
                    SAMPLE$ (2 levels)
                            A, B

     Kolmogorov-Smirnov Two Sample Test results
        Maximum differences for pairs of groups
                             A          B
                    A     0.00000000
             B     0.47500002  0.00000000
              Two-sided probabilities
                             A          B
                          A     .
                    B     0.09138048
```

< test statistic matches

< probability doesn't exactly match!