Modeling and Simulation
Supervisor: Dr. Shaista Rais
DCS-UOK

Kolmogorov-Smirnov test statistic

$$D_n = sup_x \left[|F_n(x) - F_0(x)| \right]$$

is used for testing the null hypothesis that the cumulative distribution function $F(x)$ equals some hypothesized distribution function $F_0(x)$, that is, $H_0 : F(x) = F_0(x)$, against all of the possible alternative hypotheses $H_A : F(x) \neq F_0(x)$. That is, D_n is the least upper bound of all pointwise differences $|F_n(x) - F_0(x)|$.

Justification

The bottom line is that the Kolmogorov-Smirnov statistic makes sense, because as the sample size *n* approaches infinity, the empirical distribution function $F_n(x)$ converges, with probability 1 and uniformly in *x*, to the theoretical distribution function $F(x)$. Therefore, if there is, at any point *x*, a large difference between the empirical distribution $F_n(x)$ and the hypothesized distribution $F_0(x)$, it would suggest that the empirical distribution $F_n(x)$ does not equal the hypothesized distribution $F_0(x)$. Therefore, we reject the null hypothesis:

$$H_0 : F(x) = F_0(x)$$

if D_n is too large.

Now, how do we know that $F_n(x)$ converges, with probability 1 and uniformly in *x*, to the theoretical distribution function $F(x)$? Well, unfortunately, we don't have the tools in this course to officially prove it, but we can at least do a bit of a hand-waving argument.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size *n* from a continuous distribution $F(x)$. Then, if we consider a fixed *x*, then $W = F_n(x)$ can be thought of as a random variable that takes on possible values $0, 1/n, 2/n, \ldots, 1$. Now:

- *nW* = 1, if and only if exactly 1 observation is less than or equal to *x*, and *n*−1 observations are greater than *x*
- *nW* = 2, if and only if exactly 2 observations are less than or equal to *x*, and *n*−2 observations are greater than *x*
- and in general...
- *nW* = *k*, if and only if exactly *k* observations are less than or equal to *x*, and *n*−*k* observations are greater than *x*

If we treat a success as an observation being less than or equal to *x*, then the probability of success is:

$$P(X_i \leq x) = F(x)$$

Do you see where this is going? Well, because $X_1, X_2, \ldots, X_n$ are independent random variables, the random variable *nW* is a binomial random variable with *n* trials and probability of success *p* = *F*(x). Therefore:

$$P\left(W = \frac{k}{n}\right) = P(nW = k) = \binom{n}{k}[F(x)]^k[1 - F(x)]^{n-k}$$

And, the expected value and variance of *nW* are:

$E(nW) = np = nF(x)$ and $Var(nW) = np(1-p) = n[F(x)][1 - F(x)]$

respectively. Therefore, the expected value and variance of *W* are:

$E(W) = n\frac{F(x)}{n} = F(x)$ and $Var(W) = \frac{n[(F(x)][1 - F(x)]}{n^2} = \frac{[(F(x)][1 - F(x)]}{n}$

We're very close now. We just now need to recognize that as *n* approaches infinity, the variance of *W*, that is, the variance of $F_n(x)$ approaches 0. That means that as *n* approaches infinity the empirical distribution $F_n(x)$ approaches its mean $F(x)$. And, that's why the argument for rejecting the null hypothesis if there is, at any point *x*, a large difference between the empirical distribution $F_n(x)$ and the hypothesized distribution $F_0(x)$. Not a mathematically rigorous argument, but an argument nonetheless!

Notice that the Kolmogorov-Smirnov (KS) test statistic is the supremum over *all* real x---a very large set of numbers! How then can we possibly hope to compute it? Well, fortunately, we don't have to check it at every real number but only at the sample values, since they are the only points at which the supremum can occur. Here's why:

First the easy case. If $x \geq y_n$, then $F_n(x) = 1$, and the largest difference between $F_n(x)$ and $F_0(x)$ occurs at y_n. Why? Because $F_0(x)$ can never exceed 1 and will only get closer for larger x by the monotonicity of distribution functions. So, we can record the value $F_n(y_n) - F_0(y_n) = 1 - F_0(y_n)$ and safely know that no other value $x \geq y_n$ needs to be checked.

The case where $x < y_1$ is a little trickier. Here, $F_n(x) = 0$, and the largest difference between $F_n(x)$ and $F_0(x)$ would occur at the largest possible x in this range for a reason similar to that above: $F_0(x)$ can never be negative and only gets farther from 0 for larger x. The trick is that there is no largest x in this range (since x is strictly less than y_1), and we instead have to consider lefthand limits. Since $F_0(x)$ is continuous, its limit at y_1 is simply $F_0(y_1)$. However, the lefthand limit of $F_n(y_1)$ is 0. So, the value we record is $F_0(y_1) - 0 = F_0(y_1)$, and ignore checking any other value $x < y_1$.

Finally, the general case $y_{k-1} \leq x < y_k$ is a combination of the two above. If $F_0(x) < F_n(x)$, then $F_0(y_{k-1}) \leq F_0(x) < F_n(x) = F_n(y_{k-1})$, so that $F_n(y_{k-1}) - F_0(y_{k-1})$ is at least as large as $F_n(x) - F_0(x)$ (so we don't even have to check those x values). If, however, $F_0(x) > F_n(x)$, then the largest difference will occur at the lefthand limits at y_k. Again, the continuity of F_0 allows us to use $F_0(y_k)$ here, while the lefthand limit of $F_n(y_k)$ is actually $F_n(y_{k-1})$. So, the value to record is $F_0(y_k) - F_n(y_{k-1})$, and we may disregard the other x values.

Modeling and Simulation
Supervisor: Dr. Shaista Rais
DCS-UOK

Rule for computing the KS test statistic:

For each ordered observation y_k compute the differences

$|F_n(y_k) - F_0(y_k)|$ and $|F_n(y_{k-1}) - F_0(y_k)|$.

The largest of these is the KS test statistic.

Is there any evidence to suggest that the data were not randomly sampled from a Uniform(0, 2) distribution?

Answer

The probability density function of a Uniform(0, 2) random variable *X*, say, is:

$f(x) = \frac{1}{2}$

for 0 < *x* < 2. Therefore, the probability that *X* is less than or equal to *x* is:

$P(X \leq x) = \int_0^x \frac{1}{2} dt = \frac{1}{2}x$

for 0 < *x* < 2, and we are interested in testing:

- the null hypothesis $H_0 : F(x) = F_0(x)$ against
- the alternative hypothesis $H_A : F(x) \neq F_0(x)$

where *F(x)* is the (unknown) cdf from which our data were sampled, and

$$F_0(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{1}{2}x, & \text{for } 0 \leq x < 2 \\ 1, & \text{for } x \geq 2 \end{cases}$$

Now, in working towards calculating d_n, we first need to order the eight data points so that $y_1 \leq \cdots \leq y_8$. The table below provides all the necessary values for finding the KS test statistic. Note that the empirical cdf satisfies $F_n(y_k) = k/8$, for $k = 0, 1, \ldots, 8$.

k	y_k	$F_n(y_{k-1})$	$F_0(y_k)$	$F_n(y_k)$	$\lvert F_n(y_{k-1}) - F_0(y_k)\rvert$	$\lvert F_0(y_k) - F_n(y_k)\rvert$
1	0.26	0.000	0.130	0.125	0.130	0.005
2	0.33	0.125	0.165	0.250	0.040	0.085
3	0.55	0.250	0.275	0.375	0.025	0.100
4	0.77	0.375	0.385	0.500	0.010	0.115
5	1.18	0.500	0.590	0.625	0.090	0.035
6	1.41	0.625	0.705	0.750	0.080	0.045
7	1.46	0.750	0.730	0.875	0.020	**0.145**
8	1.97	0.875	0.985	1.000	0.090	0.015

The last two columns represent all the differences we need to check. The largest of these is $d_8 = 0.145$. From the table below with $\alpha = 0.05$, the critical value is 0.46. So, we can not reject the claim that the data were sampled from Uniform(0,2).

$$D_n = \sup_x[|F_n(x) - F_0(x)|]$$
$$\alpha = 1 - P(D_n \leq d)$$

	α			
n	0.20	0.10	0.05	0.01
1	0.90	0.95	0.98	0.99
2	0.68	0.78	0.84	0.93
3	0.56	0.64	0.71	0.83
4	0.49	0.56	0.62	0.73
5	0.45	0.51	0.56	0.67
6	0.41	0.47	0.52	0.62
7	0.38	0.44	0.49	0.58
8	0.36	0.41	0.46	0.54
9	0.34	0.39	0.43	0.51
10	0.32	0.37	0.41	0.49

You might recall that the appropriateness of the *t*-statistic for testing the value of a population mean μ depends on the data being normally distributed. Therefore, one of the most common applications of the Kolmogorov-Smirnov test is to see if a set of data does follow a normal distribution. Let's take a look at an example.

Modeling and Simulation
Supervisor: Dr. Shaista Rais
DCS-UOK

Each person in a random sample of *n* = 10 employees was asked about *X*, the daily time wasted at work doing non-work activities, such as surfing the internet and emailing friends. The resulting data, in minutes, are as follows:

```
108 112 117 130 111 131 113 113 105 128
```

Is it okay to assume that these data come from a normal distribution with mean 120 and standard deviation 10?

Answer

We are interested in testing the null hypothesis, $H_0 : X$ is normally distributed with mean 120 and standard deviation 10, against the alternative hypothesis, H_A: X is not normally distributed with mean 120 and standard deviation 10. Now, in working towards calculating d_n, we again first need to order the ten data points so that $y_1 = 105$, $y_2 = 108$, etc. Then, we need to calculate the value of the hypothesized distribution function $F_0(y_k)$ at each of the values of y_k. The standard normal table can help us do this. The probability that X is less than or equal to 105, for example, equals the probability that Z is less than or equal to -1.5:

$$F_0(y_1) = P(X \leq 105) = P\left(Z \leq \tfrac{105-120}{10}\right) = P(Z \leq -1.5) = .0668.$$

The table below summarizes the relevant quantities for finding the KS test statistic. Note that the empirical cdf satisfies $F_n(y_k) = k/10$, except at $k = 5$ because of the tie: $y_5 = y_6 = 113$.

k	y_k	$F_n(y_{k-1})$	$F_0(y_k)$	$F_n(y_k)$	$\lvert F_n(y_{k-1}) - F_0(y_k)\rvert$	$\lvert F_0(y_k) - F_n(y_k)\rvert$
1	105	0.0	0.0668	0.1	0.0668	0.0332
2	108	0.1	0.1151	0.2	0.0151	0.0849
3	111	0.2	0.1841	0.3	0.0159	0.1159
4	112	0.3	0.2119	0.4	0.0881	0.1881
5	113	0.4	0.2420	0.6	0.1580	**0.3580**
6	113	0.6	0.2420	0.6	**0.3580**	**0.3580**
7	117	0.6	0.3821	0.7	0.2179	0.3179
8	128	0.7	0.7881	0.8	0.0881	0.0119
9	130	0.8	0.8413	0.9	0.0413	0.0587
10	131	0.9	0.8643	1.0	0.0357	0.1357

Modeling and Simulation
Supervisor: Dr. Shaista Rais
DCS-UOK

KOLMOGOROV SMIRNOV TEST FOR UNIFORM DISTRIBUTION WHERE F(x) = (1/2)x						
k	yk	Fn(yk-1)	Fo(yk)	Fn(yk)	|Fn(yk-1)-Fo(yk)|	|Fo(yk) - Fn(yk)|
1	0.26	0	0.13	0.125	0.13	0.005
2	0.33	0.125	0.165	0.25	0.04	0.085
3	0.55	0.25	0.275	0.375	0.025	0.1
4	0.77	0.375	0.385	0.5	0.01	0.115
5	1.18	0.5	0.59	0.625	0.09	0.035
6	1.41	0.625	0.705	0.75	0.08	0.045
7	1.46	0.75	0.73	0.875	0.02	**0.145**
8	1.97	0.875	0.985	1	0.11	0.015
		0.125=1/8	1/2*c3	b3/8		**Dn = 0.145**
		0.25=2/8				

Modeling and Simulation
Supervisor: Dr. Shaista Rais
DCS-UOK

$n \backslash \alpha$	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	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$