Activity 10 KS Test: Stat 463/563

Dr. Islam

Date: Dec 01, 2022

Last Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ First Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# One-Sample Kolmogorov-Smirnov (KS) test

Given a sample from an unknown and continuous cumulative distribution function (CDF) , the -test determines if the sample comes from a hypothesized distribution function . To test this, -test compares the empirical cumulative distribution function (ecdf) with the hypothetical distribution function.

The null and alternaive hypotheses for one-sample -test are specified as follows:

against the alternative

where is a specified CDF.

The test statistic for above test is given by

Reject the null hypothesis at significance level if , where is the critical value defined by

for a given constant

determined by the level of significance For example, given a significance level ,

and therefore, the critical value for a given sample size is

One can implement this test in R using the **ks.test()**, which provides the value of the test statistic and -value.

**Activity 10.1**

The data below refers to a sample from an unknown continuous distribution function (CDF) .

32.2 32.3 33.1 33.2 33.3 34.5 35.2 35.3 36.5 36.8 37.0 37.6

Test if the sample comes from a normal distribution with mean 35 and standard deviation 2. In order to perform this test, answer the following questions:

1. Specify null and alternative hypotheses for the test.
2. Write out the test statistic for the test.
3. Compute the value of the test statistic.
4. Find the critical value of the test at 5% level of significance.
5. What is your conclusion about the test at 5% level of significance.
6. Apply **ks.test()** available in R and report value of the test statitic and -value of the test. What is conclusion about the test at 5% level of significance using the -value.
7. Plot the ecdf of the sample, along with the ecdf of the sample for the hypothetical distribution.

**Solution**

1. and where is the CDF distribution.
2. The test statistic for above test is given by

x=scan(text="32.2 32.3 33.1 33.2 33.3 34.5 35.2 35.3 36.5 36.8 37.0 37.6")  
n=length(x)  
y=sort(x)  
#(c): Method 1  
Fn=rank(y)/n  
F0=round(pnorm(y,mean=35,sd=2),digits=4)  
Dn=max(abs(Fn-F0))  
# (c) Method 2  
ef=ecdf(x)  
Fn2=ef(x)  
Dn2=max(abs(Fn2-F0))  
c(Dn,Dn2)

[1] 0.2189667 0.2189667

#(d)  
cv=1.36/sqrt(n)  
cv

[1] 0.3925982

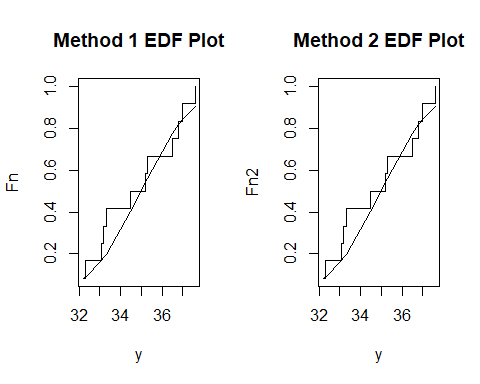
#(e) Bec Dn<cv, we accept the null hypothesis at 5% level of significance.  
c(Dn,cv)

[1] 0.2189667 0.3925982

#(f)   
ks.test(x,'pnorm',35,2)

Exact one-sample Kolmogorov-Smirnov test  
  
data: x  
D = 0.219, p-value = 0.5413  
alternative hypothesis: two-sided

#(g) plot of ecdfs  
par(mfcol=c(1,2))  
  
plot(y,Fn, type="s", main="Method 1 EDF Plot")  
lines(y,F0)  
#Or,  
plot(y,Fn2, type="s", main="Method 2 EDF Plot")  
lines(y,F0)



# Two-sample KS Test

Let and be two independent samples from populations with unknown CDFs and . Let and be corresponding EDFs. We wish to test the null hypothesis

against the alternative

The test statistic for above test is given by

For large samples, the null hypothesis is rejected at level if , where

is the critical value for the test. Note that is a constant which depends of the level of significance and is given by

For large and two sided alternative at 5% level of significance (i.e., , the cv for the test is given by

**Activity 10.2**

Two samples x and y below

x: 16 81 128 131 161 168 217 227 238 242 313 362 379 381 387 395 410 415 449 465  
y: 55 63 75 109 136 190 243 266 281 299 310 360 361 370 383 388 408

comes from two continuous but unknown CDFs and , respectively. Test if the two samples come from an indentical distribution at 5% level of significance by answering the following questions:

1. Specify null and alternative hypotheses for the test.
2. Write out the test statistic for the test.
3. Compute the value of the test statistic.
4. Find the critical value of the test at 5% level of significance.
5. What is your conclusion about the test at 5% level of significance.
6. Apply **ks.test()** available in R and report value of the test statitic and -value of the test. What is conclusion about the test at 5% level of significance using the -value.
7. Plot the ecdf of the two given samples.

**Solution**

1. against where and are unknwon CDFs of two distributions the two samples come from.
2. The test statistic is , where and are the ecdfs of two given samples evaluated at .

#x=scan(text="16 81 128 131 161 168 217 227 238 242 313 362 379 381 387 395 410 415 449 465")  
#y=scan(text="55 63 75 109 136 190 243 266 281 299 310 360 361 370 383 388 408")  
x=scan(text="387 381 395 168 81 379 238 415 242 161 131 410 128 227 313 465 362 16 449 217")  
y=scan(text="360 136 243 55 388 370 190 281 75 299 408 63 266 310 361 383 109")  
m=length(x)  
n=length(y)  
z=sort(c(x,y))  
# (c): Approach 1  
Fm=c()   
Gn=c()  
for (i in z){  
Fm=c(Fm, mean(x<=i))  
Gn=c(Gn, mean(y<=i))  
}  
Dmn=max(abs(Fm-Gn))  
# (c): Approach 2  
f2<-ecdf(x)  
g2<-ecdf(y)  
Fm2<-f2(z)  
Gn2<-g2(z)  
Dmn2=max(abs(Fm2-Gn2))  
c(Dmn, Dmn2)

[1] 0.2235294 0.2235294

# (d) CV is  
cv=1.36\*sqrt((m+n)/(m\*n))  
# (e) Since Dmn<cv, accept the null hypothesis.  
# (f) ks.test for computing test statistic and p-value  
ks.test(x,y,alt="t")

Warning in ks.test.default(x, y, alt = "t"): Parameter(s) alt ignored

Exact two-sample Kolmogorov-Smirnov test  
  
data: x and y  
D = 0.22353, p-value = 0.6502  
alternative hypothesis: two-sided

# (g) plots two ecdfs (using Approaches 1 & 2)  
par(mfcol=c(1,2))  
plot(z,Fm, type="s", main="Approch 1 EDF Plot")  
lines(z,Gn,type="s")  
  
plot(z,Fm2, type="s", main="Approch 2 EDF Plot")  
lines(z,Gn2,type="s")

