Chapter 3

Kolmogorov-Smirnov Tests

There are many situations where experimenters need to know what is the distribution of the population of their interest. For example, if they want to use a parametric test it is often assumed that the population under investigation is normal. In this chapter we consider Kolmogorov-Smirnov tests for verifying that a sample comes from a population with some known distribution and also that two populations have the same distribution.

3.1 The one-sample test

Let $x_1, \ldots, x_m$ be observations on continuous i.i.d. r.vs $X_1, \ldots, X_m$ with a c.d.f. F. We want to test the hypothesis

$$H_0 : F(x) = F_0(x) \quad \text{for all } x, \tag{3.1}$$

where F_0 is a known c.d.f.

The Kolmogorov-Smirnov test statistic D_n is defined by

$$D_n = \sup_{x \in R} |\hat{F}(x) - F_0(x)|, \tag{3.2}$$

where $\hat{F}$ is an empirical cumulative distribution defined as

$$\hat{F}(x) = \frac{\#(i : x_i \leq x)}{n}. \tag{3.3}$$

Note that supremum (3.2) must occur at one of the observed values x_i or to the left of x_i.

The null distribution of the statistic D_n can be obtained by simulation or, for large samples, using the Kolmogorov-Smirnov's distribution function.

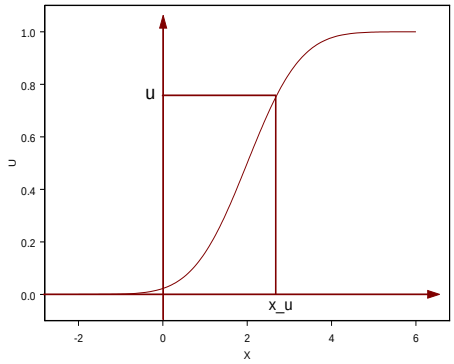

Figure 3.1: Continuous Cumulative Distribution Function

3.1.1 Simulation of the null distribution

We may approximate the null distribution of D_n by simulation. For this we use the standard uniform random variable.

Lemma 3.1
Let X be a continuous r.v. with a c.d.f. F and let $U = F(X)$. Then

$$U \sim Uniform[0, 1].$$

Proof
Let $u \in [0, 1]$. X is continuous so $\exists_{x_u \in R} F(x_u) = u$. See Figure (3.1).

Now, $F(u) = P(U < u) = P(F(X) < F(x_u)) = P(X < x_u) = F(x_u) = u$

So, $F(u) = u$ and r.v. U is uniform on $[0, 1]$.

□

To perform the simulation of D_n do the following

- generate random samples of size n from the standard uniform distribution $U[0, 1]$ with the c.d.f. $F(u) = u$,
- find maximum absolute difference between $F(u)$ and the empirical distribution $\hat{F}(u)$ for the generated sample
- repeat this N times to get the approximate distribution of D_n

Small Example
Five independent weighings of a standard weight (in $gm \times 10^{-6}$) give the following discrepancies from the supposed true weight:

$$-1.2,\ \ 0.2,\ \ -0.6,\ \ 0.8,\ \ -1.0.$$

Are the discrepancies sampled from $N(0,1)$?
We set the null hypothesis as $H_0 : F(x) = F_0(x)$ where $F_0(x) = \Phi(x)$, i.e., it is the c.d.f. of a standard normal r.v. X. To calculate the value of the test function (3.2) we need the empirical c.d.f. for the data and also the values of Φ at the data points.

The empirical c.d.f.

$$\hat{F}(x) = \begin{cases} 0 & for \quad x < -1.2 \\ 0.2 & for \quad -1.2 \le x < -1.0 \\ 0.4 & for \quad -1.0 \le x < -0.6 \\ 0.6 & for \quad -0.6 \le x < 0.2 \\ 0.8 & for \quad 0.2 \le x < 0.8 \\ 1 & for \quad x \ge 0.8 \end{cases}$$

Calculations:

x	$\hat{F}(x)$	$\Phi(x)$	$\lvert\hat{F}(x) - \Phi(x)\rvert$
-1.2^-	0	0.115	0.115
-1.2	0.2	0.115	0.085
-1.0^-	0.2	0.159	0.041
-1.0	0.4	0.159	0.241
-0.6^-	0.4	0.274	0.126
-0.6	0.6	0.274	**0.326**
$+0.2^-$	0.6	0.580	0.020
$+0.2$	0.8	0.580	0.220
$+0.8^-$	0.8	0.788	0.012
$+0.8$	1	0.788	0.212

Hence, the observed value of D_n, say d_0, is $d_0 = 0.326$

What is the null distribution of D_n?

We have $n = 5$. Suppose we have randomly generated the following five values from the standard uniform distribution:

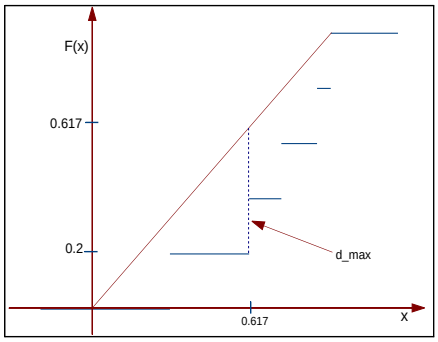

Figure 3.2: Calculating d_{max}

0.8830 0.6170 0.7431 0.9368 0.3070

D_5 gets value $d_{max} = |0.6170 - 0.2^{-}|$. See Figure 3.2

Another random sample of 5 uniform r.vs will give another value d_{max}. Repeating this procedure we simulate a set of values for D_5. Then, having the approximate null distribution of the test statistic we may calculate the $p - value$ of the observed d_0.

Below is a simulated distribution of D_5 using a GenStat program

```
     - 0.18  18 ****
0.18 - 0.24 107 **********************
0.24 - 0.30 220 *********************************************
0.30 - 0.36 231 ***********************************************
0.36 - 0.42 169 **********************************
0.42 - 0.48 119 ************************
0.48 - 0.54  80 ****************
0.54 - 0.60  35 *******
0.60 - 0.66  16 ***
0.66 -        5 *
```

This shows that $d_0 = 0.326$ is well in the middle of the distribution and so the data do not contradict the null hypothesis that the discrepancies are normally distributed with zero mean and variance equal to one.

3.1.2 Kolmogorov-Smirnov's approximation of the null distribution

The approximation is given by the following theorem.

Theorem 3.1 *Let F_0 be a continuous c.d.f., and let $X_1, \ldots, X_n$ be a sequence of i.i.d. r.vs with the c.d.f. F_0. Then*

1. *The null distribution of D_n does not depend on F_0; it depends only on n.*
2. *If $n \to \infty$ the distribution of $\sqrt{n}D_n$ is asymptotically Kolmogorov's distribution with the c.d.f.*

$$Q(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}, \tag{3.4}$$

that is

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq x) = Q(x).$$

Example: Fire Occurrences
A natural reserve in Australia had 15 fires from the beginning of this year. The fires occurred on the following days of the year: 4, 18, 32, 37, 56, 64, 78, 89, 104, 134, 154, 178, 190, 220, 256. A researcher claims that the time between the occurrences of fire in the reserve, say X, follows an exponential distribution, i.e., $X \sim Exp(\lambda)$, where $\lambda = 0.009$. Is the claim justified?

Calculations will be shown during the lectures.

3.2 The two-sample test

Let

$x_1, \ldots, x_m$ be observations on i.i.d. r.vs $X_1, \ldots, X_m$ with a c.d.f. F_1,

$y_1, \ldots, y_n$ be observations on i.i.d. r.vs $Y_1, \ldots, Y_n$ with a c.d.f. F_2.

We are interested in testing the null hypothesis of the form

$H_0 : F_1(x) = F_2(x)$ for all x

against the alternative

$H_0 : F_1(x) \neq F_2(x)$

Theorem 3.2 *Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be i.i.d. r.vs with a common continuous c.d.f. and let $\hat{F}_1$ and $\hat{F}_2$ be empirical c.d.fs of $X's$ and $Y's$, respectively. Furthermore, let*

$$D_{m,n} = \sup_t |\hat{F}_1(t) - \hat{F}_2(t)|. \tag{3.5}$$

Then we have

$$\lim_{m,n\to\infty} P\left(\sqrt{\frac{mn}{m+n}} \leq t\right) = Q(t), \tag{3.6}$$

where $Q(t)$ is given by (3.4)

Hence, the function $D_{m,n}$ may serve as a test statistic for our null hypothesis. It has asymptotic Kolmogorov-Smirnov's distribution.

NOTES

- The Kolmogorov-Smirnov (K-S) two-sample test is an alternative to the MWW test.
- The MWW test is more powerful when H_1 is the location shift. The K-S test has reasonable power against a range of alternative hypotheses.
- For small samples we may simulate the null distribution of $D_{m,n}$ applying standard uniform distribution $U[0, 1]$.

Learning the mechanics example will be given during the lectures.