Lecture 28

28.1 Kolmogorov-Smirnov test.

Suppose that we have an i.i.d. sample $X_1, \ldots, X_n$ with some unknown distribution $\mathbb{P}$ and we would like to test the hypothesis that $\mathbb{P}$ is equal to a particular distribution $\mathbb{P}_0$, i.e. decide between the following hypoheses:

$$\begin{cases} H_1: & \mathbb{P} = \mathbb{P}_0 \\ H_2: & \text{otherwise} \end{cases}$$

We considered this problem before when we talked about goodness-of-fit test for continuous distribution but, in order to use Pearson's theorem and chi-square test, we discretized the distribution and considered a weaker derivative hypothesis. We will now consider a different test due to Kolmogorov and Smirnov that avoids this discretization and in a sense is more consistent.

Let us denote by $F(x) = \mathbb{P}(X_1 \le x)$ a cumulative distribution function and consider what is called an empirical distribution function:

$$F_n(x) = \mathbb{P}_n(X \le x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

that is simply the proportion of the sample points below level x. For any fixed point $x \in \mathbb{R}$ the law of large numbers gives that

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) \to \mathbb{E}I(X_1 \le x) = \mathbb{P}(X_1 \le x) = F(x),$$

i.e. the proportion of the sample in the set $(-\infty, x]$ approximates the probability of this set.

It is easy to show from here that this approximation holds uniformly over all $x \in \mathbb{R}$:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0$$

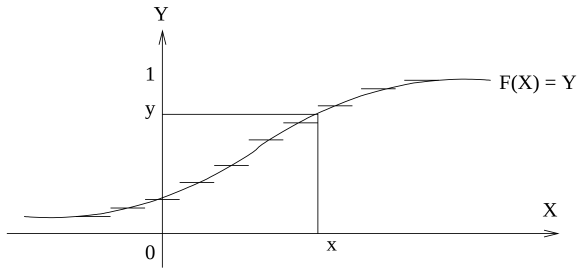

Figure 28.1: C.d.f. and empirical d.f.

i.e. the largest difference between F_n and F goes to 0 in probability. The key observation in the Kolmogorov-Smirnov test is that the distribution of this supremum does not depend on the distribution $\mathbb{P}$ of the sample.

Theorem 1. *The distribution of* $\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|$ *does not depend on* F.

Proof. For simplicity, let us assume that F is continuous, i.e. the distribution is continuous. Let us define the inverse of F by

$$F^{-1}(y)=\min\{x: F(x)\geq y\}.$$

Then making the change of variables $y=F(x)$ or $x=F^{-1}(y)$ we can write

$$\mathbb{P}(\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\leq t)=\mathbb{P}(\sup_{0\leq y\leq 1}|F_n(F^{-1}(y))-y|\leq t).$$

Using the definition of the empirical d.f. F_n we can write

$$F_n(F^{-1}(y))=\frac{1}{n}\sum_{i=1}^n I(X_i\leq F^{-1}(y))=\frac{1}{n}\sum_{i=1}^n I(F(X_i)\leq y)$$

and, therefore,

$$\mathbb{P}(\sup_{0\leq y\leq 1}|F_n(F^{-1}(y))-y|\leq t)=\mathbb{P}\Big(\sup_{0\leq y\leq 1}\Big|\frac{1}{n}\sum_{i=1}^n I(F(X_i)\leq y)-y\Big|\leq t\Big).$$

The distribution of $F(X_i)$ is uniform on the interval $[0,1]$ because the c.d.f. of $F(X_1)$ is

$$\mathbb{P}(F(X_1)\leq t)=\mathbb{P}(X_1\leq F^{-1}(t))=F(F^{-1}(t))=t.$$

Therefore, the random variables

$$U_i=F(X_i)\text{ for }i\leq n$$

are independent and have uniform distribution on $[0,1]$ and, combining with the above, we proved that

$$\mathbb{P}(\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\leq t)=\mathbb{P}\Big(\sup_{0\leq y\leq 1}\Big|\frac{1}{n}\sum_{i=1}^{n}I(U_i\leq y)-y\Big|\leq t\Big)$$

which is clearly independent of F.

□

Next, we will formulate the main result on which the KS test is based. First of all, let us note that for a fixed x the CLT implies that

$$\sqrt{n}(F_n(x)-F(x))\to N\Big(0,F(x)(1-F(x))\Big)$$

because $F(x)(1-F(x))$ is the variance of $I(X_1\leq x)$. If turns out that if we consider

$$\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|$$

it will also converge to some distribution.

Theorem 2. *We have,*

$$\mathbb{P}(\sqrt{n}sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\leq t)\to H(t)=1-2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t}$$

where $H(t)$ is the c.d.f. of Kolmogorov-Smirnov distribution.

If we formulate our hypotheses in terms of cumulative distribution functions:

$$\begin{cases} H_1: & F=F_0 \text{ for a given} F_0 \\ H_2: & \text{otherwise} \end{cases}$$

then based on Theorems 1 and 2 the Kolmogorov-Smirnov test is formulated as follows:

$$\delta=\begin{cases} H_1: & D_n\leq c \\ H_2: & D_n>c \end{cases}$$

where

$$D_n=\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F_0(x)|$$

and the threshold c depends on the level of significance α and can be found from the condition

$$\alpha=\mathbb{P}(\delta\neq H_1|H_1)=\mathbb{P}(D_n\geq c|H_1).$$

In Theorem 1 we showed that the distribution of D_n does not depend on the unknown distribution F and, therefore, it can tabulated. However, the distribution of D_n

depends on n so one needs to use advanced tables that contain the table for the sample size n of interest. Another way to find c, especially when the sample size is large, is to use Theorem 2 which tells that the distribution of D_n can be approximated by the Kolmogorov-Smirnov distribution and, therefore,

$$\alpha = \mathbb{P}(D_n \geq c|H_1) \approx 1 - H(c).$$

and we can use the table for H to find c.

To explain why Kolmogorov-Smirnov test makes sense let us imagine that the first hypothesis fails and H_2 holds which means that $F \neq F_0$.

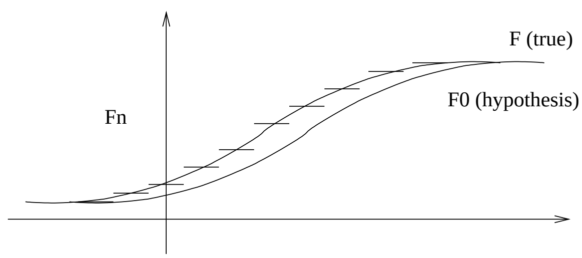

Figure 28.2: The case when $F \neq F_0$.

Since F is the true c.d.f. of the data, by law of large numbers the empirical d.f. F_n will converge to F as shown in figure 28.2 and as a result it will not approximate F_0, i.e. for large n we will have

$$\sup_x |F_n(x) - F_0(x)| > \delta$$

for small enough δ. Multiplying this by $\sqrt{n}$ will give that

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| > \sqrt{n}\delta.$$

If H_1 fails then $D_n > \sqrt{n}\delta \to +\infty$ as $n \to \infty$. Therefore, it seems natural to reject H_1 when D_n becomes too large which is exactly what happens in KS test.

$\square$

Example. Let us consider a sample of size 10:

$$0.58, 0.42, 0.52, 0.33, 0.43, 0.23, 0.58, 0.76, 0.53, 0.64$$

and let us test the hypothesis that the distribution of the sample is uniform on $[0, 1]$:

$$\begin{cases} H_1 : & F(x) = F_0(x) = x \\ H_2 : & \text{otherwise} \end{cases}$$

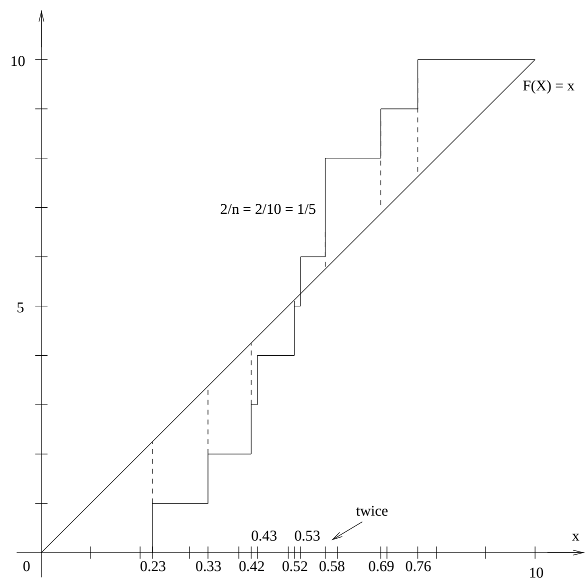

Figure 28.3: F_n and F_0 in the example.

The figure 28.3 shows the c.d.f. F_0 and empirical d.f. $F_n(x)$.

To compute D_n we notice that the largest difference between $F_0(x)$ and $F_n(x)$ is achieved either before or after one of the jumps, i.e.

$$\sup_{0\le x\le 1} |F_n(x) - F(x)| = \max_{1\le i\le n} \begin{cases} |F_n(X_i^-) - F(X_i)| & \text{- before the } i\text{th jump} \\ |F_n(X_i) - F(X_i)| & \text{- after the } i\text{th jump} \end{cases}$$

Writing these differences for our data we get

before the jump	after the jump
$\lvert 0 - 0.23\rvert$	$\lvert 0.1 - 0.23\rvert$
$\lvert 0.1 - 0.33\rvert$	$\lvert 0.2 - 0.33\rvert$
$\lvert 0.2 - 0.42\rvert$	$\lvert 0.3 - 0.42\rvert$
$\lvert 0.3 - 0.43\rvert$	$\lvert 0.4 - 0.43\rvert$
$\cdots$	

The largest value will be achieved at $|0.9 - 0.64| = 0.26$ and, therefore,

$$D_n = \sqrt{n} \sup_{0\le x\le 1} |F_n(x) - x| = \sqrt{10} \times 0.26 = 0.82.$$

If we take the level of significance $\alpha = 0.05$ then

$$1 - H(c) = 0.05 \Rightarrow c = 1.35$$

and according to KS test

$$\delta = \begin{cases} H_1 : & D_n \leq 1.35 \\ H_2 : & D_n > 1.35 \end{cases}$$

we accept the null hypothesis H_1 since $D_n = 0.82 < c = 1.35$.