Kolmogorov-Smirnov Test

Samahriti Mukherjee , Aytijhya Saha

8 August 2021

Let $X_1, X_2, .., X_n$ be a sample from distribution function F and $Y_1, Y_2, ..., Y_m$ be another sample from distribution function G. We want to test $H_0 : F = G$ vs $H_1 : F \neq G$.
K-S Test Statistic = $sup_{x\in\mathbb{R}}|F_n\hat{(x)} - G_m\hat{(x)}|$
We shall reject the null hypothesis for large values of test-statistic.
We have to check that we can calculate the K-S test Statistic by evaluating $\hat{F_n}$ and $\hat{G_m}$ only at finitely many points. Relate this thing with a Random Walk problem.

We know that

$$\hat{F_n}(x) = \frac{\sum_{i=1}^{n} 1_{[X_i \leq x]}}{n}$$

and

$$\hat{G_m}(x) = \frac{\sum_{i=1}^{m} 1_{[X_i \leq x]}}{m}$$

$\hat{F_n}(x)$ is a step function discontinuous only at distinct elements of the set . Also, $\hat{G_m}(x)$ is a step function discontinuous only at distinct elements of the set $\{Y_1, Y_2, .., Y_m\}$. So, $\hat{F_n}(x) - \hat{G_m}(x)$ is a step function discontinuous only at distinct elements of the set $\{X_1, X_2, .., X_n, Y_1, Y_2, .., Y_m\} = A$, say. So as we can calculate the difference $\hat{F_n}(x) - \hat{G_m}(x)$ everywhere in the domain $x \in \mathbb{R}$ by evaluating the difference only at distinct elements of A, we can calculate the K-S test Statistic by evaluating $\hat{F_n}$ and $\hat{G_m}$ only at finitely many points.(Proved)

Let us now assume, m=n.
Now we arrange the elements of A in non-decreasing order
Let $z_1 \leq z_2 \leq .... \leq z_{2m}$ be the elements of A. For an interpretation in terms of paths,we write ϵ_p = +1 or -1 according as z_j equals to X_i, for some i or Y_i, for some i.

Claim: $|\hat{F_m}(t) - \hat{G_m}(t)| > c$ for some t if and only if $|s_k| > cm$ for some k , and $c > 0$.

Let, $t \in [z_k, z_{k+1})$, for some k.

Now when $|\hat{F_m}(t) - \hat{G_m}(t)| > c$, then $|s_k| = m|\hat{F_m}(t) - \hat{G_m}(t)| > cm$

When given that, $|s_k| > cm$ for some k,take $t = z_k$, then, $|\hat{F_m}(t) - \hat{G_m}(t)| > c$.(proved)

Let, our test statistic, $sup_{x\in\mathbb{R}}|\hat{F_n}(x) - \hat{G_m}(x)| \leq M$, then
$\hat{F_n}(x) - \hat{G_m}(x) \leq M \forall x$
$\iff$ $|s_k| \leq Mm$ for all $k \in \{1, 2, .., m\}$, as if $|s_k| > Mm$ for some k, then $|\hat{F_n}(x) - \hat{G_m}(x)| > Mm$ by our claim which we proved earlier, but M is the supremum of $\hat{F_n}(x) - \hat{G_m}(x)$
$\iff$ $sup\{|s_k| : k = 1, .., m\} \leq mM$.

Thus,we can relate the test with random walk problem.