Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is used to test if an empirical_cumulative_distribution follows a particular distribution.

Glivenko-Cantelli Theorem (Fundamental theorem of statistics)

Let $F(t)$ be the true CDF of $X_1, \ldots, X_n \stackrel{iid}{\sim} X$. Let $F_n(t)$ be the empirical cdf of $X_1, \ldots, X_n$. Then,

$$\sup_{t\in \mathbb{R}}|F_n(t)-F(t)|\xrightarrow[n\rightarrow\infty]{a.s.}0$$

This tells us that as the number of samples increases, the empirical and true CDFs will converge for all values of t.

Asymptotic normality

$$\sqrt{n}(F_n(t)-F(t))\xrightarrow[n\to\infty]{(d)}\mathcal{N}(0,F(t)(1-F(t)))$$

Donsker's Theorem

If F is continuous, then

$$\sqrt{n}\sup_{t\in\mathbb{R}}|F_n(t)-F(t)|\xrightarrow[n\to\infty]{(d)}\sup_{0<t'<1}|\mathbb{B}(t')|$$

$\mathbb{B}$ is the a Brownian bridge on $[0,1]$, which is defined:

$$\mathbb{B}(t') \sim \mathcal{N}(0,t')$$

It is called a Brownian bridge because the values at $t'=0$ and $t'=1$ are pinned at 0.

Hypothesis test setup

Let $X_1, \ldots, X_n$ be i.i.d. random variables and follow the cdf F. Let F^0 be a continuous cdf. This test that tells us whether those values follow the cdf F^0

$$H_0: F = F^0$$ $$H_1: F \neq F^0$$

Let F_n be the empirical cdf of the sample $X_1, \ldots, X_n$. If $F=F^0$, then $F_n(t)\approx F^0(t)$ for $t\in [0,1]$.

The test statistic is:

$$T_n = \sup_{t\in\mathbb{R}}|F_n(t)-F^0(t)|$$

By Donsker's theorem, if H_0 is true, then,

$$\sqrt{n}T_n \xrightarrow [n\to \infty]{(d)} Z$$

where Z is the supremum of a Brownian bridge from $t'=0$ to $t'=1$.

The Kolmogorov-Smirnov with asymptotic level α is defined as:

$$\delta_\alpha=\mathbb{1}\{T_n > \frac{q_\alpha}{\sqrt{n}}\}$$

where q_α is the ($1-\alpha$) quantile of Z.

And the p-value is:

$$p = \mathbb{P}[Z > T_n]$$

Calculating the KS test statistic

Let $X_{(i)}$ be the ith smallest sample. ($X_{(1)}\leq X_{(2)}\leq \ldots X_{(n)}$). Then, the formula for the test statistic T_n becomes:

$$T_n=\max_i\{\max(|\frac{i-1}{n}-F^0(X_{(i)})|, |\frac{i}{n}-F^0(X_{(i)})|)\}$$

This formula checks all of the samples (discontinuities in the empirical cdf) and finds the maximum distance between the empirical cdf and the cdf we are checking against (F^0).

This test statistic is a pivotal statistic because it does not depend on the distribution of X_is. In other words, this test statistic is general for all KS tests, not specific to any one.