# 4.7 Kolmogorov-Smirnov test

## Overview

The Kolmogorov-Smirnov test is a non-parametric test for comparing two distributions of a continuous variable. Rejection of the null hypothesis indicates that the two distributions differ from one another in some way (location, dispersion, skewness, etc.). The evolution of the test can be traced in the work of Kolmogorov (1933) , Kolmogorov (1941) , Smirnov (1939a) and Smirnov (1939b) . See also Darling (1957) for a detailed synopsis.

## The null hypothesis

There are essentially two versions of the test in common usage: one is a goodness-of-fit test, designed to compare the distribution of a sampled random variable with some known distribution. The other is to compare the distributions of two sampled random variables ('the two-sample problem' *sensu* Darling (1957) ). The Kolmogorov-Smirnov test in PRIMER implements this latter (two-sample) test.

- Let $X_1, X_2, \ldots, X_{n_1}$ be a set of $n_1$ observations of independent random variables ('sample 1' or 'group 1') that each have the same continuous distribution function, $U(x)= \text{Pr} \lbrace{ X_i < x \rbrace}$.
- Similarly, we let $Y_1, Y_2, \ldots, Y_{n_2}$ be a set of $n_2$ observations of independent random variables ('sample 2' or 'group 2') that each have the same continuous distribution function, $V(x)= \text{Pr} \lbrace{ Y_i < x \rbrace}$.

The null hypothesis for the Kolmogorov-Smirnov test is that these two groups of samples come from the same distribution, i.e. $$ \text{H}_ 0 \text{:} \hspace{0.2cm} U(x) = V(x) $$

## Description of the test statistic

Let $\hat{F}_ {n_1}(x)$ be the empirical distribution function for group 1. Specifically, $\hat{F}_ {n_1}(x)$ is the proportion of the $X_i$, $i = 1, \ldots, n_1$, that are less than $x$. Thus, if $X_i = 20$, then $\hat{F}_ {n_1}(X_i)$ is the proportion of the $n_1$ values in group 1 that are less than 20. This will be equal to one minus the proportion of $n_1$ values in group 1 that are greater than or equal to 20. Similarly, we can let $\hat{G}_ {n_2}(x)$ be the empirical distribution function for group 2, defined in the same way but for that group.

Once these two empirical distributions have been calculated, the Kolmogorov-Smirnov test-statistic is defined as

$$ D = \sup_{-\infty < x < \infty} |\hat{F}_ {n_1}(x) - \hat{G}_ {n_2}(x)| $$

Thus, $D$ is the supremum of the absolute values of differences calculated between the two empirical distribution functions. In essence, the test-statistic here captures the largest possible difference that is observable between the two empirical distributions for any value of $x$.

## Calculating a *p*-value

There are tabled values for $D$ that can be calculated under certain conditions, but in PRIMER we simply generate a distribution for $D$ under the null hypothesis directly and empirically by assuming only *exchangeability* between the two groups. For the permutation test, the $(n_1 + n_2)$ values are permuted randomly across the two groups (preserving each of the individual group's original sample sizes, $n_1$ and $n_2$, respectively), and the test statistic is calculated for the permuted data as $D^\pi$. We can repeat this randomisation procedure a large number of times (e.g., $n_\text{perm}$ = 9999), to obtain a permutation distribution of $D^\pi$ under a true null hypothesis.

By comparing our observed value (obtained with the original ordering of the data, $D_\text{obs}$) with the distribution of $D^\pi$, we obtain a direct empirical estimate of the probability associated with the null hypothesis. Specifically, if we let $D_k^\pi$ be the value of $D^\pi$ obtained for the $k$th permutation ($k = 1, \ldots, n_\text{perm}$), the *p*-value is calculated as the proportion of $D_k^\pi$ values that equal or exceed $D_\text{obs}$:

$$ p = \frac{ \sum_{k=1}^{n_\text{perm}} (\text{I}(D_k^\pi \geq D_\text{obs} ) + 1 )}{(n_\text{perm} + 1)} $$

with $\text{I}(\textit{expression} ) = 1$ if $\textit{expression}$ is true and zero otherwise.(Note that the "+1" in the numerator and denominator here acknowledge and include $D_\text{obs}$ as a member of the permutation distribution; the original order is one possible ordering of the data, after all!)

---

Revision #68
Created 19 November 2024 03:06:07 by Marti
Updated 29 November 2025 16:46:35 by Marti