A Kolmogorov-Smirnov type test for two inter-dependent random variables

Tommy Liu*
February 28, 2018

Abstract
Consider \( n \) iid random variables, where \( \xi_1, \ldots, \xi_n \) are \( n \) realisations of a random variable \( \xi \) and \( \zeta_1, \ldots, \zeta_n \) are \( n \) realisations of a random variable \( \zeta \). The distribution of each realisation of \( \xi \), that is the distribution of one \( \xi_i \), depends on the value of the corresponding \( \zeta_i \), that is the probability \( P(\xi_i \leq x) = F(x, \zeta_i) \). We develop a statistical test to see if the \( \xi_1, \ldots, \xi_n \) are distributed according to the distribution function \( F(x, \zeta_i) \). We call this new statistical test the condition Kolmogorov-Smirnov test.

Contents
1 Introduction 1
2 Kolmogorov-Smirnov Test 2
3 Conditional Kolmogorov-Smirnov Test 2
4 Conclusion 4
References 4

1 Introduction
Random variables often occur in experiments. It is common to have \( n \) observations of a quantity, which gives rise to \( n \) identically and independently distributed random variables \( \xi_1, \ldots, \xi_n \), where each \( \xi_i \) (where \( i = 1, \ldots, n \)) are different realisation of the same random variable \( \xi \). Analysis is then done to test whether the measured \( \xi_i \) follow a particular distribution. This problem was famously answered by Kolmogorov in [1].

In this paper we consider a variant of this problem. In a physical experiment, \( n \) measurements of two quantities \( \xi \) and \( \zeta \) are taken, resulting in \( n \) iid pairs of random variables, with \( \xi_1, \ldots, \xi_n \) being \( n \) realisations of \( \xi \) and \( \zeta_1, \ldots, \zeta_n \) being \( n \) realisations of \( \zeta \). A distribution of \( \zeta \) is not known, but the distribution of one realisation of \( \xi \), that is \( \xi_i \) depends on the corresponding \( \zeta_i \).

We want to test whether the pairs of \((\xi_i, \zeta_i)\) we collected follow a particular joint distribution. The answer we offer is very similar to the original Kolmogorov-Smirnov test.

*University of Reading tommy.liu.academic@gmail.com
2 Kolmogorov-Smirnov Test

First we recall well known results about the Kolmogorov-Smirnov statistic and the Kolmogorov-Smirnov test [1]. Let

\[ \xi_1, \xi_2, \ldots, \xi_n \]

be \( n \) independently and identically distributed real random variables. Each \( \xi_i \) is distributed with CDF \( F(\cdot) \) as in

\[ P(\xi_i \leq x) = F(x) = \int_{-\infty}^{x} f(s) \, ds \]

where \( f(\cdot) \) is the PDF. Define a function by \( F_n(\cdot) \) by

\[ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(\xi_i) \]

where \( 1_A \) is the indicator function for a set \( A \). Thus \( F_n(\cdot) \) is the empirical CDF. Consider the distance between the real and the empirical CDF through the supremum metric on the space of real functions

\[ D_n = \|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(\xi_i) - F(x) \right| \]

where \( D_n \) is called the Kolmogorov-Smirnov statistic or KS statistic. We define what we mean by the null hypothesis.

**Definition 2.1.** Let \( \xi_1, \xi_2, \ldots, \xi_n \) be \( n \) real random variables. The null hypothesis is that each \( \xi_i \) is independently distributed with CDF \( F(\cdot) \).

We want to know how large or small \( D_n \) needs to be before deciding whether to reject the null hypothesis. The following Theorem offers a remarkable answer to this problem.

**Theorem 2.2.** Suppose the null hypothesis is true, then the distribution of \( D_n \) depends only on \( n \).

Notice that \( D_n \) is in itself a real random variable. The PDF and CDF of \( D_n \) is a function of \( n \) only, and will be the same whatever \( F(\cdot) \) is. This distribution is called the KS distribution and tables are available up to \( n = 100 \). There is a Theorem which describes the asymptotic behaviour of the KS distribution.

**Theorem 2.3.** In the limit \( n \to \infty \), \( \sqrt{n}D_n \) is asymptotically Kolmogorov distributed with the CDF

\[ Q(x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2} \]

that is to say

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

3 Conditional Kolmogorov-Smirnov Test

The conditional Kolmogorov-Smirnov test was first developed in the thesis [2]. Let \( \zeta_1, \zeta_2, \ldots, \zeta_n \) be \( n \) iid real random variables. They are \( n \) empirical observations of a random variable \( \zeta \). Now suppose that each of the \( \xi_1, \xi_2, \ldots, \xi_n \) is conditioned and dependent on the corresponding \( \zeta_1, \zeta_2, \ldots, \zeta_n \). The conditional CDF \( F(\cdot, \cdot) \) is

\[ P(\xi \leq x \mid \zeta_i) = F_{\zeta_i}(x) = \int_{-\infty}^{x} f(s, \zeta_i) \, ds \]
But \( \xi_1, \xi_2, \ldots, \xi_n \) are empirical measurements of the same random variable \( \xi \). The CDF for \( \xi \) is

\[
P(\xi \leq x) = F(x) = \int_{-\infty}^{x} \int_{-\infty}^{u=+\infty} f(s, u) m(u) du \, ds
\]

where \( m(\cdot) \) is the PDF for \( \zeta \), that is

\[
P(\zeta \in A) = \int_A m(s) ds
\]

In our context we have the problem that the random variables are not identically distributed under the null hypothesis. The \( \xi_1, \xi_2, \ldots, \xi_n \) and \( \zeta_1, \zeta_2, \ldots, \zeta_n \) are obtained experimentally and \( F_{\zeta_i}(\xi_i) \) can be calculated but a PDF for \( \zeta_i \), that is \( m(\cdot) \), has no easy expression. We still want to perform a statistics test that is similar to the KS test even in such situations where the distribution \( m(\cdot) \) of \( \zeta \) is unknown. First we define what we call the total null hypothesis and the conditional null hypothesis.

**Definition 3.1.** Let \( \xi_1, \xi_2, \ldots, \xi_n \) be \( n \) empirical observations of a random variable \( \xi \). The total null hypothesis is that \( \xi \) is distributed with the CDF \( F(\cdot) \). The conditional null is that each \( \xi_i \) is distributed with the conditional CDF \( F_{\zeta_i}(\cdot) \).

A new statistical test is developed, which is similar to the KS test.

**Theorem 3.2.** Suppose the conditional null hypothesis is true. Let \( F_{\zeta_i}(\cdot) \) be continuous. Let \( S_n \) be the statistic given by

\[
S_n = \sup_{x \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{[0,x]}(F_{\zeta_i}(\xi_i)) - x \right|
\]

then \( S_n \) is KS distributed.

**Proof.** Denote

\[
Y_i = F_{\zeta_i}(\xi_i)
\]

which means

\[
P(Y_i \leq x) = P(F_{\zeta_i}(\xi_i) \leq x) = P(\xi_i \leq F_{\zeta_i}^{-1}(x)) = F_{\zeta_i}(F_{\zeta_i}^{-1}(x)) = x
\]

and \( 0 \leq Y_i \leq 1 \), so \( Y_i \) is uniformly distributed on \([0,1]\). Note that \( F_{\zeta_i}(\cdot) \) is a function of one variable only. Let

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[0,x]}(Y_i) = \frac{1}{n} \sum_{i=1}^{n} 1_{[0,x]}(F_{\zeta_i}(\xi))
\]

where \( F_n(\cdot) \) is the empirical CDF of a uniformly distributed random variable, computed using \( n \) observations. The statistic \( S_n \) is the supremum metric

\[
S_n = \|F_n - x\|_{\infty} = \sup_{x \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{[0,x]}(F_{\zeta_i}(\xi_i)) - x \right|
\]

So clearly \( S_n \) is KS distributed. \( \square \)
We call $S_n$ the conditional KS statistic. Compare this to the original KS statistic, which under the assumption of the total null hypothesis can be rewritten as

$$D_n = \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x]}(\xi_i) - F(x) \right| = \sup_{x \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{[0,x]}(F(\xi_i)) - x \right|$$

When both the total and conditional hypothesis are true $D_n$ and $S_n$ are KS distributed, that is

$$P(D_n \in A) = P(S_n \in A) \quad \text{and} \quad P(D_n \leq x) = P(S_n \leq x)$$

The subtlety here is that $D_n$ and $S_n$ are different objects, yet they have the same distribution. $D_n$ is KS distributed under the total null hypothesis, whereas $S_n$ is KS distributed under the conditional null hypothesis. This can be explained in another way. We have $n$ experimental observations of a random variable $\xi$ denoted by $\xi_1, \xi_2, \ldots, \xi_n$ and each are conditioned on observations of another random variable $\zeta$ denoted by $\zeta_1, \zeta_2, \ldots, \zeta_n$. The $D_n$ is KS distributed if the random variable $\xi$ is distributed by CDF $F(\cdot)$, but the $S_n$ is KS distributed if each $\xi_i$ is conditionally distributed by the CDF $F_{\zeta_i}(\cdot)$.

4 Conclusion

An example of the conditional Kolmogorov-Smirnov test reliably working with experimental data is studied in the paper [3], which is also the first time of it being implemented.

References

[1] A. N. Kolmogorov, “Sulla Determinazione Empirica di una Legge di Distribuzione,” Giornale dell’Istituto Italiano degli Attuari, vol. 4, pp. 83–91, 1933.

[2] T. Liu, Stochastic Resonance for a Model with Two Pathways. PhD thesis, University of Reading, 2016. Available on arXiv.

[3] T. Kuna, V. Lucarini, and T. Liu, “A comparative study of stochastic resonance for a model with two pathways by escape times, linear response, invariant measures and the conditional kolmogorov-smirnov test,” 2018. Available on arXiv.