A distance metric between histograms through the the Kolmogorov-Smirnov test statistic: specification, measures reference and example uses

Renato Fabbri[1, a)]
São Carlos Institute of Physics, University of São Paulo (IFSC/USP), PO Box 369, 13560-970, São Carlos, SP, Brazil

(Dated: 31 October 2015)

This document presents reference values for a distance metric derived from the Kolmogorov-Smirnov test statistic $D_{F,F'}$. Each measure of $D_{F,F'}$ is a distance between two histograms, which is normalized by the number of observations in each sample to yield the c statistic, which can be mapped to p-values, i.e. values for which high enough levels of significance α implies the rejection of the null hypothesis. Benchmarks for the implementation are delivered by comparing samples from known distributions. Pattern examples in real data enables further insight in the robustness and power of c.

PACS numbers: 05.10-a,
Keywords: Kolmogorov-Smirnov test, statistic, benchmark, distance measure, histogram

CONTENTS

I. INTRODUCTION

Be F and F' two empirical cumulative distributions, where n and n' are the number of observations on each sample. The two-sample Kolmogorov-Smirnov test rejects the null hypothesis that the histograms are the outcome of the same underlying distribution if:

$$D_{F,F'} > c(\alpha)\sqrt{\frac{n+n'}{nn'}} \tag{1}$$

where $D_{F,F'} = sup_x[F - F']$ as in Figure 1 and $c(\alpha)$ is related to the level of significance α by:

α	0.1	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

a) http://ifsc.usp.br/~fabbri/; Electronic mail: fabbri@usp.br

If distributions are drawn from empirical data, $D_{F,F'}$ is given as are n and n'. All terms in equation 1 are positive and $c(\alpha)$ can be isolated:

$$c(\alpha) < \frac{D_{n,n'}}{\sqrt{\frac{n+n'}{nn'}}} = c \tag{2}$$

When c is high, low values of α favor rejecting the null hypothesis. In fact, c can be normalized to yield p-values.

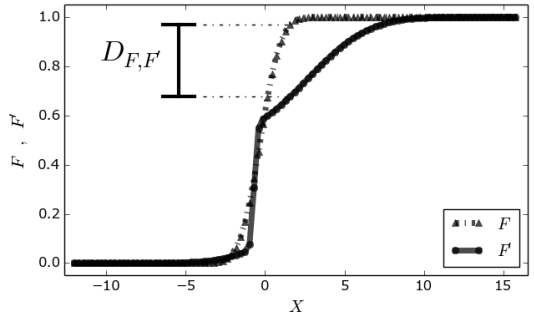

FIG. 1. The Kolmogorov-Smirnov statistic $D_{F,F'}$: the maximum difference between two cumulative distribution functions.

High values of c favor rejecting the null hypothesis. For example, if the significance level is $\alpha = 0.01$, then c greater than 1.7 implies the rejection of the null hypothesis and suggests that F and F' are outcomes of different distributions. Of core importance in this study is to regard the c statistic as a measure of distance between both distributions[1]. The main contribution of the following sections is the explicit display of reference values of c from which one might derive knowledge from measures or even from a single value of c.

A. Philosophical and technological note

Difference and equivalence is of central role in human cognition, philosophy and science. This fact is so deeply recognized that thinkers often reduce thought to classifications, e.g. through the mathematical concept of equivalence classes[2]. Histograms are very immediate and informative roughly wherever there is a phenomenon of interest which can yield measurements. This present document should enable conclusions to be drawn about the equivalence (and difference) of the processes underlying sets of measurements for a very broad range of phenomena. The following tables validate the mathematical framework and the software implementation.

B. Document outline

Section II exposes reference values drawn from simulations. Section III exemplifies the use of the c statistic to make sense of phenomena. Section IV holds final remarks with directions to software and data.

II. REFERENCES THROUGH SIMULATIONS

Values of c are given for simulations involving normal, uniform, Weibull and power function distributions. The rendering of this article is automated to ease changes in the settings with which the results are reported. The number of comparisons is $N_c = 1000$, each with the sample size of $n = n' = 100000$. Each histogram have $N_b = 300$ equally spaced bins.

A. When the null hypothesis is true

If the null hypothesis is true, the number of rejections of the null hypothesis ($c > c(\alpha)$) in N_c comparisons should not exceed αN_c. To verify this, let $C = \{c_i\}$ be a set of c measures, and $C(\alpha) = \{c : c > c(\alpha)\}$. Be $|C(\alpha)|$ the cardinality of $C(\alpha)$, i.e. the number of comparisons in which the two-sample Kolmogorov-Smirnov test rejects the null hypothesis for a given α. This section reports that $|C(\alpha)|$ very rarely exceeds αN_c, for all probability distributions and settings. Also important are that $c > c(\alpha)$ in many cases and that the tabulated α values are also good estimates of the upper limit of the frequency of such an event.

αN_c	α	$c(\alpha)$	$\lvert C_1(\alpha)\rvert$	$\lvert C_2(\alpha)\rvert$	$\lvert C_3(\alpha)\rvert$
10.0	0.100	1.22	7	10	8
5.0	0.050	1.36	5	3	4
2.5	0.025	1.48	3	1	3
1.0	0.010	1.63	0	1	3
0.5	0.005	1.73	0	0	1
0.1	0.001	1.95	0	0	0

TABLE I. The theoretical maximum number αN_c of rejections of the null hypothesis for significance levels α. The c_1 values were calculated using simulations of normal distributions with $\mu = 0$ and $\sigma = 1$. The c_2 values were calculated using simulations of normal distributions with $\mu = 3$ and $\sigma = 2$. The c_3 values were calculated using simulations of normal distributions with $\mu = 6$ and $\sigma = 3$. Over all N_c comparisons, $\mu(c_1) = 0.8218$ and $\sigma(c_1) = 0.2515$, $\mu(c_2) = 0.7811$ and $\sigma(c_2) = 0.2748$, $\mu(c_3) = 0.8365$ and $\sigma(c_3) = 0.2523$.

αN_c	α	$c(\alpha)$	$\lvert C_1(\alpha)\rvert$	$\lvert C_2(\alpha)\rvert$	$\lvert C_3(\alpha)\rvert$
10.0	0.100	1.22	9	8	8
5.0	0.050	1.36	4	3	3
2.5	0.025	1.48	1	2	2
1.0	0.010	1.63	0	0	2
0.5	0.005	1.73	0	0	1
0.1	0.001	1.95	0	0	0

TABLE II. The theoretical maximum number αN_c of rejections of the null hypothesis for critical values of α. The c_1 values were calculated using simulations of uniform distributions within $[0, 1)$. The c_2 values were calculated using simulations of uniform distributions within $[2, 6)$. The c_3 values were calculated using simulations of uniform distributions with $\mu = 4$ and $\sigma = 10$. Over all N_c comparisons, $\mu(c_1) = 0.8616$ and $\sigma(c_1) = 0.2291$, $\mu(c_2) = 0.8345$ and $\sigma(c_2) = 0.2492$, $\mu(c_3) = 0.8086$ and $\sigma(c_3) = 0.2740$.

αN_c	α	$c(\alpha)$	$\lvert C_1(\alpha)\rvert$	$\lvert C_2(\alpha)\rvert$	$\lvert C_3(\alpha)\rvert$	$\lvert C_4(\alpha)\rvert$
10.0	0.100	1.22	0	10	6	6
5.0	0.050	1.36	0	4	2	4
2.5	0.025	1.48	0	0	1	2
1.0	0.010	1.63	0	0	1	2
0.5	0.005	1.73	0	0	0	2
0.1	0.001	1.95	0	0	0	1

TABLE III. The theoretical maximum number αN_c of rejections of the null hypothesis for critical values of α. The c_1 values were calculated using simulations of 1-parameter Weibull distributions with $a = 0.1$. The c_2 values were calculated using simulations of 1-parameter Weibull distributions with $a = 2$. The c_3 values were calculated using simulations of 1-parameter Weibull distributions with $a = 4$. Over all N_c comparisons, The N_o values of c_4 were calculated using simulations of 1-parameter Weibull distributions with $a = 6$. Over all N_c comparisons, $\mu(c_1) = 0.1073$ and $\sigma(c_1) = 0.0686$, $\mu(c_2) = 0.8285$ and $\sigma(c_2) = 0.2486$, $\mu(c_3) = 0.7891$ and $\sigma(c_3) = 0.2334$. $\mu(c_4) = 0.8063$ and $\sigma(c_4) = 0.2845$.

αN_c	α	$c(\alpha)$	$\lvert C_1(\alpha)\rvert$	$\lvert C_2(\alpha)\rvert$	$\lvert C_3(\alpha)\rvert$	$\lvert C_4(\alpha)\rvert$	$\lvert C_5(\alpha)\rvert$
10.0	0.100	1.22	6	12	12	9	7
5.0	0.050	1.36	6	3	5	4	2
2.5	0.025	1.48	2	0	1	2	1
1.0	0.010	1.63	1	0	1	1	0
0.5	0.005	1.73	1	0	1	1	0
0.1	0.001	1.95	0	0	0	0	0

TABLE IV. The theoretical maximum number αN_c of rejections of the null hypothesis for critical values of α. The c_1 values were calculated using simulations of power functions distributions with $a = 0.3$. The c_2 values were calculated using simulations of power functions distributions with $a = 1$. The c_3 values were calculated using simulations of power functions distributions with $a = 2$. The c_4 values were calculated using simulations of power functions distributions with $a = 3$. The c_5 values were calculated using simulations of power functions distributions with $a = 4$. Over all N_c comparisons, $\mu(c_1) = 0.8253$ and $\sigma(c_1) = 0.2697$, $\mu(c_2) = 0.8470$ and $\sigma(c_2) = 0.2557$, $\mu(c_3) = 0.8799$ and $\sigma(c_3) = 0.2588$. $\mu(c_4) = 0.8278$ and $\sigma(c_4) = 0.2637$. $\mu(c_5) = 0.7795$ and $\sigma(c_5) = 0.2364$.

B. When the null hypothesis if false

The null hypothesis is always false for a sufficiently small significance level α. In this section, each table holds a set comparisons between two samples: one sample is generated through a fixed distribution while the other sample is modified in each comparison. The comparison is repeated N_c times. The measures on c chosen to report the results are: the mean $\mu(c)$, the standard deviation $\sigma(c)$, the median $m(c)$, the fraction $\overline{C(\alpha)} = \frac{|C(\alpha)|}{N_c}$ of rejection of the null hypothesis given the significance level α. The null hypothesis is true in the boldface lines.

σ	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.5	3.965	0.309	3.958	3.354,3.421,3.444	4.606,4.696,4.763	1.000	1.000	1.000	1.000	1.000	1.000
0.6	3.072	0.276	3.075	2.348,2.549,2.571	3.667,3.712,3.913	1.000	1.000	1.000	1.000	1.000	1.000
0.7	2.287	0.243	2.292	1.632,1.766,1.834	2.795,2.817,2.952	1.000	1.000	1.000	1.000	0.990	0.910
0.8	1.718	0.263	1.699	1.118,1.118,1.185	2.214,2.326,2.393	0.960	0.920	0.830	0.650	0.440	0.220
0.9	1.131	0.290	1.129	0.537,0.559,0.626	1.789,1.834,1.901	0.350	0.190	0.110	0.080	0.040	0.000
1.0	**0.847**	**0.275**	**0.805**	**0.402,0.447,0.447**	**1.565,1.588,1.632**	**0.110**	**0.070**	**0.040**	**0.010**	**0.000**	**0.000**
1.1	1.023	0.252	1.006	0.537,0.581,0.626	1.655,1.699,1.722	0.210	0.090	0.040	0.030	0.000	0.000
1.2	1.450	0.292	1.465	0.648,0.693,0.760	1.945,1.990,1.990	0.760	0.650	0.480	0.260	0.180	0.020
1.3	1.792	0.250	1.766	1.163,1.163,1.297	2.303,2.393,2.437	0.980	0.960	0.900	0.790	0.560	0.250
1.4	2.179	0.276	2.180	1.543,1.610,1.677	2.728,2.817,3.086	1.000	1.000	1.000	0.980	0.970	0.790
1.5	2.542	0.288	2.527	1.923,1.923,1.923	3.153,3.466,3.578	1.000	1.000	1.000	1.000	1.000	0.970
1.6	2.863	0.258	2.862	2.303,2.370,2.393	3.354,3.511,3.533	1.000	1.000	1.000	1.000	1.000	1.000
1.7	3.174	0.259	3.164	2.683,2.706,2.706	3.712,3.824,3.846	1.000	1.000	1.000	1.000	1.000	1.000
1.8	3.424	0.287	3.399	2.885,2.907,2.907	4.092,4.114,4.226	1.000	1.000	1.000	1.000	1.000	1.000
1.9	3.743	0.258	3.712	3.198,3.265,3.309	4.360,4.494,4.562	1.000	1.000	1.000	1.000	1.000	1.000
2.0	4.027	0.287	4.014	3.421,3.511,3.533	4.696,4.696,4.740	1.000	1.000	1.000	1.000	1.000	1.000

TABLE V. Measurements of c through simulations with normal distributions. One normal distribution is fixed, with $\mu = 0$ and $\sigma = 1$, and compared agaist normal distributions with $\mu = 0$ and different values of σ.

μ	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.0	**0.838**	**0.259**	**0.805**	**0.313,0.425,0.447**	**1.453,1.476,1.789**	**0.070**	**0.040**	**0.010**	**0.010**	**0.010**	**0.000**
0.1	1.280	0.410	1.252	0.581,0.604,0.648	2.191,2.326,2.348	0.540	0.410	0.310	0.180	0.110	0.080
0.2	2.178	0.386	2.169	1.163,1.230,1.588	3.019,3.063,3.376	0.990	0.980	0.980	0.950	0.890	0.710
0.3	2.997	0.456	2.974	1.856,2.102,2.147	3.801,3.824,3.846	1.000	1.000	1.000	1.000	1.000	0.990
0.4	3.812	0.363	3.868	2.929,2.952,2.952	4.494,4.539,5.009	1.000	1.000	1.000	1.000	1.000	1.000
0.5	4.697	0.423	4.685	3.645,3.824,3.980	5.590,5.635,5.702	1.000	1.000	1.000	1.000	1.000	1.000
0.6	5.479	0.505	5.478	4.360,4.450,4.494	6.440,6.462,6.574	1.000	1.000	1.000	1.000	1.000	1.000
0.7	6.383	0.445	6.440	5.344,5.389,5.389	7.178,7.267,7.491	1.000	1.000	1.000	1.000	1.000	1.000
0.8	7.152	0.433	7.167	6.216,6.261,6.261	7.983,7.983,8.206	1.000	1.000	1.000	1.000	1.000	1.000
0.9	7.948	0.389	7.972	6.909,7.155,7.155	8.631,8.654,8.676	1.000	1.000	1.000	1.000	1.000	1.000
1.0	8.758	0.434	8.765	7.290,7.692,7.759	9.481,9.615,9.794	1.000	1.000	1.000	1.000	1.000	1.000

TABLE VI. Measurements of c through simulations with normal distributions. One normal distribution is fixed, with $\mu = 0$ and $\sigma = 1$, and compared agaist normal distributions with different values of μ and fixed $\sigma = 1$.

b	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.7	6.774	0.328	6.775	6.149,6.216,6.216	7.357,7.357,7.558	1.000	1.000	1.000	1.000	1.000	1.000
0.75	5.612	0.303	5.624	4.763,4.808,4.875	6.149,6.283,6.328	1.000	1.000	1.000	1.000	1.000	1.000
0.8	4.504	0.260	4.483	3.980,4.003,4.003	5.121,5.143,5.232	1.000	1.000	1.000	1.000	1.000	1.000
0.85	3.418	0.272	3.421	2.773,2.885,2.907	3.935,4.047,4.226	1.000	1.000	1.000	1.000	1.000	1.000
0.9	2.373	0.263	2.381	1.789,1.878,1.968	2.885,2.952,3.019	1.000	1.000	1.000	1.000	1.000	0.980
0.95	1.366	0.237	1.330	0.894,0.939,0.962	1.901,1.923,2.057	0.700	0.460	0.290	0.140	0.090	0.010
1.0	**0.806**	**0.255**	**0.783**	**0.402,0.447,0.447**	**1.453,1.543,1.744**	**0.080**	**0.030**	**0.020**	**0.010**	**0.010**	**0.000**
1.05	1.399	0.275	1.364	0.872,0.872,0.984	2.057,2.147,2.258	0.760	0.540	0.310	0.170	0.100	0.060
1.1	2.213	0.230	2.191	1.632,1.766,1.834	2.750,2.840,2.862	1.000	1.000	1.000	1.000	0.990	0.910
1.15	3.011	0.233	3.019	2.348,2.393,2.460	3.421,3.421,3.712	1.000	1.000	1.000	1.000	1.000	1.000
1.2	3.763	0.289	3.779	2.929,3.086,3.108	4.204,4.316,4.562	1.000	1.000	1.000	1.000	1.000	1.000
1.25	4.568	0.268	4.573	3.958,3.980,4.003	5.098,5.121,5.255	1.000	1.000	1.000	1.000	1.000	1.000
1.3	5.202	0.287	5.210	4.137,4.494,4.696	5.769,5.858,5.881	1.000	1.000	1.000	1.000	1.000	1.000
1.35	5.828	0.301	5.836	4.986,5.143,5.165	6.328,6.350,6.373	1.000	1.000	1.000	1.000	1.000	1.000

TABLE VII. Measurements of c through simulations with uniform distributions. One uniform distribution has the fixed domain $[0, 1)$. The other uniform distribution in each comparison is also centered around 0.5, but spread over $b = b_u - b_l$ there b_l and b_u are the lower and upper boudaries.

μ	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.5	**0.875**	**0.254**	**0.850**	**0.358,0.380,0.447**	**1.409,1.521,1.632**	**0.090**	**0.080**	**0.020**	**0.010**	**0.000**	**0.000**
0.55	1.680	0.323	1.666	1.096,1.163,1.185	2.370,2.460,2.661	0.960	0.840	0.690	0.540	0.430	0.170
0.6	2.872	0.347	2.873	2.102,2.147,2.191	3.690,3.712,3.734	1.000	1.000	1.000	1.000	1.000	1.000
0.65	3.940	0.314	3.913	3.265,3.421,3.444	4.651,4.651,4.673	1.000	1.000	1.000	1.000	1.000	1.000
0.7	5.018	0.309	5.009	4.360,4.383,4.405	5.613,5.680,5.948	1.000	1.000	1.000	1.000	1.000	1.000
0.75	6.166	0.328	6.194	5.456,5.568,5.568	6.708,6.731,7.245	1.000	1.000	1.000	1.000	1.000	1.000
0.8	7.261	0.336	7.222	6.574,6.596,6.596	8.005,8.027,8.095	1.000	1.000	1.000	1.000	1.000	1.000
0.85	8.308	0.296	8.318	7.759,7.782,7.782	8.922,9.078,9.078	1.000	1.000	1.000	1.000	1.000	1.000
0.9	9.452	0.333	9.425	8.765,8.877,8.922	10.018,10.129,10.152	1.000	1.000	1.000	1.000	1.000	1.000
0.95	10.582	0.335	10.577	9.883,9.928,9.951	11.225,11.247,11.382	1.000	1.000	1.000	1.000	1.000	1.000
1.0	11.620	0.368	11.594	10.845,10.912,10.912	12.343,12.455,12.969	1.000	1.000	1.000	1.000	1.000	1.000
1.05	12.734	0.312	12.768	11.784,11.963,12.008	13.327,13.349,13.595	1.000	1.000	1.000	1.000	1.000	1.000
1.1	13.879	0.345	13.819	13.103,13.126,13.238	14.579,14.602,14.736	1.000	1.000	1.000	1.000	1.000	1.000

TABLE VIII. Measurements of c through simulations with uniform distributions. One uniform distribution has the fixed domain $[0, 1)$. The other uniform distribution in each comparison have varied mean values but always spread over a fixed $b = b_u - b_l$ there b_l and b_u are the lower and upper boudaries.

a	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.01	0.034	0.019	0.022	0.022,0.022,0.022	0.089,0.089,0.112	0.000	0.000	0.000	0.000	0.000	0.000
0.1	0.396	0.205	0.402	0.045,0.067,0.089	0.783,0.805,1.073	0.000	0.000	0.000	0.000	0.000	0.000
0.3	5.205	0.556	5.277	2.817,3.421,3.466	6.015,6.194,6.216	1.000	1.000	1.000	1.000	1.000	1.000
0.5	6.093	0.413	6.060	5.009,5.322,5.344	7.044,7.066,7.088	1.000	1.000	1.000	1.000	1.000	1.000
0.7	4.473	0.387	4.494	3.533,3.622,3.734	5.322,5.389,5.434	1.000	1.000	1.000	1.000	1.000	1.000
0.9	3.161	0.387	3.097	2.460,2.504,2.527	3.935,4.092,4.360	1.000	1.000	1.000	1.000	1.000	1.000
1.1	2.088	0.317	2.113	1.409,1.521,1.521	2.773,2.862,2.952	1.000	1.000	0.990	0.910	0.860	0.660
1.3	1.241	0.321	1.252	0.626,0.671,0.738	1.901,2.080,2.571	0.520	0.320	0.170	0.120	0.060	0.020
1.5	**0.774**	**0.233**	**0.760**	**0.358,0.380,0.380**	**1.342,1.364,1.453**	**0.060**	**0.020**	**0.000**	**0.000**	**0.000**	**0.000**
1.7	1.134	0.304	1.096	0.581,0.604,0.671	1.789,1.968,2.258	0.350	0.190	0.120	0.090	0.030	0.020
1.9	1.768	0.313	1.755	1.185,1.185,1.230	2.460,2.594,2.773	0.980	0.940	0.790	0.630	0.530	0.270
2.1	2.343	0.392	2.337	1.588,1.722,1.744	3.220,3.220,3.421	1.000	1.000	1.000	0.990	0.980	0.800
2.3	2.796	0.394	2.750	2.012,2.080,2.102	3.645,3.645,3.935	1.000	1.000	1.000	1.000	1.000	1.000
2.5	3.200	0.359	3.220	2.370,2.370,2.549	3.846,4.003,4.070	1.000	1.000	1.000	1.000	1.000	1.000
2.7	3.714	0.413	3.656	2.907,2.929,2.952	4.651,4.897,4.942	1.000	1.000	1.000	1.000	1.000	1.000
2.9	3.991	0.351	4.025	3.130,3.175,3.198	4.606,4.629,4.718	1.000	1.000	1.000	1.000	1.000	1.000

TABLE IX. Measurements of c through simulations with 1-parameter Weibull distributions. One Weibull distribution has the fixed shape parameter $a = 1.5$. The other Weibull distribution in each comparison has varied values of a.

a	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.7	6.319	0.463	6.328	5.210,5.232,5.411	7.178,7.200,7.312	1.000	1.000	1.000	1.000	1.000	1.000
0.9	4.480	0.471	4.528	3.511,3.555,3.555	5.322,5.367,5.747	1.000	1.000	1.000	1.000	1.000	1.000
1.1	2.746	0.429	2.773	1.632,1.856,1.901	3.555,3.846,4.003	1.000	1.000	1.000	1.000	0.990	0.960
1.3	1.562	0.413	1.599	0.671,0.693,0.783	2.303,2.348,2.482	0.800	0.660	0.540	0.490	0.350	0.200
1.5	**0.810**	**0.247**	**0.771**	**0.402,0.425,0.425**	**1.364,1.364,1.632**	**0.090**	**0.030**	**0.010**	**0.010**	**0.000**	**0.000**
1.7	1.439	0.442	1.409	0.559,0.626,0.626	2.370,2.415,2.661	0.670	0.570	0.420	0.320	0.240	0.150
1.9	2.242	0.446	2.258	1.297,1.386,1.409	3.130,3.376,3.533	1.000	0.990	0.940	0.910	0.870	0.720
2.1	2.985	0.488	2.952	1.699,1.811,1.945	4.003,4.137,4.293	1.000	1.000	1.000	1.000	0.990	0.970
2.3	3.820	0.453	3.857	2.795,2.907,2.929	4.696,4.740,4.852	1.000	1.000	1.000	1.000	1.000	1.000
2.5	4.474	0.470	4.450	3.533,3.555,3.622	5.568,5.568,5.747	1.000	1.000	1.000	1.000	1.000	1.000

TABLE X. Measurements of c through simulations with power function distributions. One power distribution has the fixed exponent parameter $1 - a = 2.5$. The other power function distribution in each comparison has varied values of a.

III. EXAMPLE USES IN EMPIRICAL DATA

This section presents immediate results drawn from the statistic c when observed in real samples. The sample choices are arbitrary.

A. Text

This section exemplifies the use of c in the detection of similarity between texts. Each text X was divided in two halves $X1$ and $X2$. The set of known English words were considered as were the set of stopwords (words with reduced meaning such as prepositions and articles). Only the number of letters in each words was measured. Three approaches were chosen: 1) the text was partitioned into 1000 pieces of equal number of characters, the mean of the word size of each piece is an element of the sample; 2) the text was partitioned into 1000 pieces of equal number of characters, the standard deviation of the word size is an element of the sample; 3) each word size is an element of the sample. This last case yields a discrete probability distribution, which was approximated as a continuous variable and gave the greatest sensibility to text differences. The overall result is the same: smaller differences between parts of the same text. Notice that the c is often high within a same book.

label	description	chars	tokens	sentences	$\mu(kw)$	$\sigma(kw)$	$\mu(sw)$	$\sigma(sw)$
H,H1,H2	Hamlet by Shakespeare	162881	37360	3106	3.549	1.762	2.721	1.011
B,B1,B2	King James Version of the Holly Bible	4332554	1010654	30103	3.745	1.711	2.927	1.044
M,M1,M2	Moby Dick by Herman Melville	1242990	260819	10059	4.105	2.184	2.847	1.096
E,E1,E2	Esaú e Jacó from Machado de Assis	355706	88472	3822	2.186	1.376	1.486	0.502

TABLE XI. General description of the texts used to exemplify the use of the c statistic. Individual values of number of characters, tokens, sentences give context. Mean and standard deviation of the size of known words kw and of the stopwords st are used in next tables. Numbers in the labels indicate first and second half of the corresponding text in the next tables.

	H	**H1**	**H2**	**B**	**B1**	**B2**	**M**	**M1**	**M2**	**E**	**E1**	**E2**
H	0.000	2.996	3.198	11.426	9.749	11.985	14.691	13.864	13.819	20.298	19.528	19.141
H1	2.996	0.000	0.537	12.298	10.644	12.634	13.998	13.349	13.461	18.461	17.416	17.061
H2	3.198	0.537	0.000	12.544	11.493	13.618	15.093	14.378	14.266	18.504	17.528	17.173
B	11.426	12.298	12.544	0.000	3.824	2.750	15.339	13.685	13.998	22.249	21.799	21.757
B1	9.749	10.644	11.493	3.824	0.000	5.791	15.339	13.707	14.154	22.137	21.663	21.645
B2	11.985	12.634	13.618	2.750	5.791	0.000	13.886	12.164	12.097	22.271	21.866	21.779
M	14.691	13.998	15.093	15.339	15.339	13.886	0.000	1.766	1.744	22.249	21.753	21.757
M1	13.864	13.349	14.378	13.685	13.707	12.164	1.766	0.000	0.872	22.136	21.686	21.623
M2	13.819	13.461	14.266	13.998	14.154	12.097	1.744	0.872	0.000	22.114	21.685	21.667
E	20.298	18.461	18.504	22.249	22.137	22.271	22.249	22.136	22.114	0.000	2.971	2.133
E1	19.528	17.416	17.528	21.799	21.663	21.866	21.753	21.686	21.685	2.971	0.000	1.163
E2	19.141	17.061	17.173	21.757	21.645	21.779	21.757	21.623	21.667	2.133	1.163	0.000

TABLE XII. Values of c for histograms drawn from mean of the sizes of the known words.

	H	**H1**	**H2**	**B**	**B1**	**B2**	**M**	**M1**	**M2**	**E**	**E1**	**E2**
H	0.000	2.817	3.690	6.507	4.875	7.066	12.634	11.091	11.113	8.545	11.318	10.420
H1	2.817	0.000	0.984	8.721	7.267	9.705	13.394	12.500	12.298	5.961	8.838	8.072
H2	3.690	0.984	0.000	10.062	8.296	10.465	14.311	13.170	12.992	4.860	8.348	7.759
B	6.507	8.721	10.062	0.000	3.801	2.571	15.720	13.752	14.043	14.620	15.983	14.982
B1	4.875	7.267	8.296	3.801	0.000	5.769	16.659	14.467	14.825	13.273	15.156	13.841
B2	7.066	9.705	10.465	2.571	5.769	0.000	14.154	12.030	12.746	15.134	16.529	15.496
M	12.634	13.394	14.311	15.720	16.659	14.154	0.000	2.258	1.856	17.964	18.285	17.821
M1	11.091	12.500	13.170	13.752	14.467	12.030	2.258	0.000	0.626	16.952	17.923	17.352
M2	11.113	12.298	12.992	14.043	14.825	12.746	1.856	0.626	0.000	16.908	17.673	17.061
E	8.545	5.961	4.860	14.620	13.273	15.134	17.964	16.952	16.908	0.000	4.389	3.843
E1	11.318	8.838	8.348	15.983	15.156	16.529	18.285	17.923	17.673	4.389	0.000	1.160
E2	10.420	8.072	7.759	14.982	13.841	15.496	17.821	17.352	17.061	3.843	1.160	0.000

TABLE XIII. Values of c' for histograms drawn from the standard deviation of the sizes of the known words.

	H	H1	H2	B	B1	B2	M	M1	M2	E	E1	E2
H	0.000	0.650	0.656	10.207	10.393	9.704	12.611	12.216	11.729	41.743	33.650	33.528
H1	0.650	0.000	1.131	8.092	8.278	7.779	8.917	8.852	8.484	34.046	29.074	28.982
H2	0.656	1.131	0.000	6.457	6.656	6.159	9.428	9.352	8.986	35.161	30.060	29.967
B	10.207	8.092	6.457	0.000	6.831	6.683	29.218	22.346	21.408	65.138	46.483	46.284
B1	10.393	8.278	6.656	6.831	0.000	11.703	30.425	24.194	23.312	64.556	46.384	46.188
B2	9.704	7.779	6.159	6.683	11.703	0.000	22.666	18.113	17.199	63.969	45.943	45.748
M	12.611	8.917	9.428	29.218	30.425	22.666	0.000	0.617	0.612	60.900	44.199	44.013
M1	12.216	8.852	9.352	22.346	24.194	18.113	0.617	0.000	1.065	58.252	43.168	42.993
M2	11.729	8.484	8.986	21.408	23.312	17.199	0.612	1.065	0.000	58.239	43.139	42.963
E	41.743	34.046	35.161	65.138	64.556	63.969	60.900	58.252	58.239	0.000	0.250	0.251
E1	33.650	29.074	30.060	46.483	46.384	45.943	44.199	43.168	43.139	0.250	0.000	0.434
E2	33.528	28.982	29.967	46.284	46.188	45.748	44.013	42.993	42.963	0.251	0.434	0.000

TABLE XIV. Values of c' for histograms drawn from the sizes of the known words.

	H	H1	H2	B	B1	B2	M	M1	M2	E	E1	E2
H	0.000	2.333	2.095	10.800	10.196	9.481	7.111	5.724	5.613	21.375	21.417	21.391
H1	2.333	0.000	0.882	11.547	10.988	10.183	8.333	7.215	7.170	19.255	19.297	19.270
H2	2.095	0.882	0.000	10.568	9.919	8.981	7.941	6.957	6.890	19.700	19.741	19.715
B	10.800	11.547	10.568	0.000	1.521	1.722	5.635	6.239	7.044	22.338	22.361	22.338
B1	10.196	10.988	9.919	1.521	0.000	1.073	4.830	5.456	6.283	22.338	22.361	22.334
B2	9.481	10.183	8.981	1.722	1.073	0.000	3.958	4.606	5.322	22.338	22.361	22.334
M	7.111	8.333	7.941	5.635	4.830	3.958	0.000	1.588	1.856	22.315	22.338	22.312
M1	5.724	7.215	6.957	6.239	5.456	4.606	1.588	0.000	1.207	22.292	22.334	22.307
M2	5.613	7.170	6.890	7.044	6.283	5.322	1.856	1.207	0.000	22.315	22.334	22.307
E	21.375	19.255	19.700	22.338	22.338	22.338	22.315	22.292	22.315	0.000	2.472	3.751
E1	21.417	19.297	19.741	22.361	22.361	22.361	22.338	22.334	22.334	2.472	0.000	1.465
E2	21.391	19.270	19.715	22.338	22.334	22.334	22.312	22.307	22.307	3.751	1.465	0.000

TABLE XV. Values of c' for histograms drawn from the mean of the sizes of the stopwords.

	H	H1	H2	B	B1	B2	M	M1	M2	E	E1	E2
H	0.000	4.216	5.241	7.759	5.568	7.290	8.587	6.865	6.127	20.861	20.880	20.827
H1	4.216	0.000	1.026	10.876	8.916	10.406	11.546	9.914	9.243	16.578	16.597	16.544
H2	5.241	1.026	0.000	11.105	8.962	10.635	11.775	10.210	9.410	15.710	15.729	15.676
B	7.759	10.876	11.105	0.000	2.907	1.431	2.929	2.862	2.706	22.315	22.334	22.281
B1	5.568	8.916	8.962	2.907	0.000	2.437	4.942	2.907	2.549	22.315	22.334	22.281
B2	7.290	10.406	10.635	1.431	2.437	0.000	2.639	1.565	1.655	22.315	22.334	22.281
M	8.587	11.546	11.775	2.929	4.942	2.639	0.000	2.169	2.661	22.338	22.334	22.281
M1	6.865	9.914	10.210	2.862	2.907	1.565	2.169	0.000	0.648	22.315	22.334	22.281
M2	6.127	9.243	9.410	2.706	2.549	1.655	2.661	0.648	0.000	22.315	22.334	22.281
E	20.861	16.578	15.710	22.315	22.315	22.315	22.338	22.315	22.315	0.000	4.870	6.237
E1	20.880	16.597	15.729	22.334	22.334	22.334	22.334	22.334	22.334	4.870	0.000	1.497
E2	20.827	16.544	15.676	22.281	22.281	22.281	22.281	22.281	22.281	6.237	1.497	0.000

TABLE XVI. Values of c' for histograms drawn from the standard deviation of the sizes of the stopwords.

	H	**H1**	**H2**	**B**	**B1**	**B2**	**M**	**M1**	**M2**	**E**	**E1**	**E2**
H	0.000	0.835	0.847	10.183	10.950	9.075	4.219	4.131	3.858	28.322	21.836	21.139
H1	0.835	0.000	1.456	8.314	8.916	7.563	4.081	4.064	3.857	24.599	19.810	19.248
H2	0.847	1.456	0.000	6.196	6.810	5.472	2.055	2.100	1.893	25.815	20.821	20.237
B	10.183	8.314	6.196	0.000	3.777	3.811	14.417	10.427	11.101	38.938	28.127	27.106
B1	10.950	8.916	6.810	3.777	0.000	6.571	15.289	11.548	12.190	39.275	28.450	27.424
B2	9.075	7.563	5.472	3.811	6.571	0.000	10.935	8.189	8.804	38.128	27.622	26.624
M	4.219	4.081	2.055	14.417	15.289	10.935	0.000	0.422	0.416	34.862	25.388	24.474
M1	4.131	4.064	2.100	10.427	11.548	8.189	0.422	0.000	0.725	34.184	25.154	24.267
M2	3.858	3.857	1.893	11.101	12.190	8.804	0.416	0.725	0.000	34.032	25.032	24.149
E	28.322	24.599	25.815	38.938	39.275	38.128	34.862	34.184	34.032	0.000	0.382	0.410
E1	21.836	19.810	20.821	28.127	28.450	27.622	25.388	25.154	25.032	0.382	0.000	0.686
E2	21.139	19.248	20.237	27.106	27.424	26.624	24.474	24.267	24.149	0.410	0.686	0.000

TABLE XVII. Values of c' for histograms drawn from sizes of the stopwords.

B. Audio

This section presents c values drawn from audio for testing the sound system of the computer. The PCM samples of the files were normalized to fit the interval $[-1, 1]$ to yield the samples labeled. The wavelet decomposition was performed with the Daubechies 8 Wavelet function. The resulting values of the c statistic reflect most of all the different types of signals analysed: PCM samples and wavelet decomposition coefficients in different leafs. Among each type of signal, the type of sound is also reflected in the measures of c, with the noise having the highest values.

label	description	events
S1	recorded 'front center'	68545.00
W1-1	first wavelet approximation	31.00
W2-1	higher wavelet leaf	17147.00
S2	recorded 'front left'	71042.00
W1-2	first wavelet approximation	32.00
W2-2	higher wavelet leaf	17771.00
S3	recorded 'rear center'	65026.00
W1-3	first wavelet approximation	30.00
W2-3	higher wavelet leaf	16267.00
S4	recorded 'rear left'	63010.00
W1-4	first wavelet approximation	30.00
W2-4	higher wavelet leaf	15763.00
S5	noise	67579.00
W1-5	first wavelet approximation	31.00
W2-5	higher wavelet leaf	16906.00

TABLE XVIII. General description of the audio data used for the c values of the next table. The recorded data events are the PCM samples normalized to fit [-1,1]. The wavelet first approximation consists of the low frequencies. The higher leaf consists of an approximation of one of the last details.

	S1	**W1-1**	**W2-1**	**S2**	**W1-2**	**W2-2**	**S3**	**W1-3**	**W2-3**	**S4**	**W1-4**	**W2-4**	**S5**	**W1-5**	**W2-5**
S1	0.000	1.788	22.327	8.142	2.062	20.747	13.854	2.054	19.074	10.855	1.968	22.221	33.542	1.702	19.829
W1-1	1.788	0.000	0.638	0.695	0.736	2.121	1.303	0.579	2.657	1.080	1.956	2.599	1.541	1.778	1.129
W2-1	22.327	0.638	0.000	17.427	1.138	2.867	30.634	0.851	2.242	25.737	2.858	4.600	42.026	2.251	29.901
S2	8.142	0.695	17.427	0.000	0.476	20.678	21.162	0.684	16.983	16.482	2.262	21.187	41.117	1.689	19.076
W1-2	2.062	0.736	1.138	0.476	0.000	3.030	0.565	0.721	2.944	0.603	1.599	1.552	0.748	1.540	0.842
W2-2	20.747	2.121	2.867	20.678	3.030	0.000	34.710	2.034	2.981	29.594	3.365	1.602	46.165	2.511	33.095
S3	13.854	1.303	30.634	21.162	0.565	34.710	0.000	1.282	32.838	7.304	1.456	35.427	26.085	1.430	24.678
W1-3	2.054	0.579	0.851	0.684	0.721	2.034	1.282	0.000	3.090	1.063	1.936	2.244	1.228	1.763	1.208
W2-3	19.074	2.657	2.242	16.983	2.944	2.981	32.838	3.090	0.000	25.000	2.901	3.176	39.066	2.258	30.109
S4	10.855	1.080	25.737	16.482	0.603	29.594	7.304	1.063	25.000	0.000	1.711	28.490	28.993	1.532	21.878
W1-4	1.968	1.956	2.858	2.262	1.599	3.365	1.456	1.936	2.901	1.711	0.000	3.411	2.062	1.377	2.648
W2-4	22.221	2.599	4.600	21.187	1.552	1.602	35.427	2.244	3.176	28.490	3.411	0.000	44.885	2.512	33.243
S5	33.542	1.541	42.026	41.117	0.748	46.165	26.085	1.228	39.066	28.993	2.062	44.885	0.000	2.305	16.793
W1-5	1.702	1.778	2.251	1.689	1.540	2.511	1.430	1.763	2.258	1.532	1.377	2.512	2.305	0.000	2.407
W2-5	19.829	1.129	29.901	19.076	0.842	33.095	24.678	1.208	30.109	21.878	2.648	33.243	16.793	2.407	0.000

TABLE XIX. Values of c for histograms drawn from sound PCM samples and wavelet leaf coefficients. The different types of the signals yield greater c values.

C. Music

This section presents measures of the c statistic drawn from the pitches of the notes of classical compositions. The results reflect music history. For example, measures of c involving Palestrina increases with the exception of Beethoven who, indeed, used modalism. The values of c related to Bach also increases along time, and the outcome of the comparison against Palestrina is only exceeded when Schönberg is reached, which reflects the non-tonal discourse of both Palestrina and Schönberg.

label	description	events
Pale	Sanctus 69 from G. P. da Palestrina	719.00
Bach1	BWV735 from J. S. Bach	236.00
Bach2	BWV648 from J. S. Bach	272.00
Moza1	K80 from W. A. Mozart	538.00
Moza2	K458 from W. A. Mozart	4218.00
Beet1	Opus 18, n1, mov. 3 from L. van Beethoven	1289.00
Beet2	Opus 132 from L. van Beethoven	17884.00
Schön	Opus 19, mov. 2 from A. Schönberg	102.00

TABLE XX. General description of the music data used for the c values of the next table. Each event is a midi value of a note pitch. Samples where chosen to reflect music history timeline. Works by the same composer were chosen among the first and last 10% of all he produced.

	Pale	**Bach1**	**Bach2**	**Moza1**	**Moza2**	**Beet1**	**Beet2**	**Schön**
Pale	0.00	1.36	1.05	1.41	2.17	1.35	1.68	1.89
Bach1	1.36	0.00	0.88	1.00	0.76	0.74	0.95	1.49
Bach2	1.05	0.88	0.00	1.73	1.22	1.25	0.83	1.27
Moza1	1.41	1.00	1.73	0.00	2.14	1.32	0.60	1.79
Moza2	2.17	0.76	1.22	2.14	0.00	1.49	3.32	1.84
Beet1	1.35	0.74	1.25	1.32	1.49	0.00	2.34	1.86
Beet2	1.68	0.95	0.83	0.60	3.32	2.34	0.00	1.92
Schön	1.89	1.49	1.27	1.79	1.84	1.86	1.92	0.00

TABLE XXI. Values of c for histograms drawn from the pitches of classical compositions.

D. OS status

This last example expose the statistic c for samples drawn from the operational system of my laptop. The patterns are less neat than on last examples, but many conclusions can still be reached. The memory used by the most consuming processes compose the samples which present the highest values of c. Lowest values of c are related to RAM usage. Again, the type of samples are mandatory: they might all be identified by the values of c found in comparison to other samples, with the exception of the RAM memory.

label	description	events
cpu1	workload of the most active processor	888.00
cpu2	workload of the second most active processor	422.00
cpu3	workload of the third most active processor	1046.00
mem	RAM use in kB	557.00
p1	workload use of most consuming process	1197.00
m1	RAM use of most consuming process	1197.00
p2	workload use of second most consuming process	1197.00
m2	RAM use of second most consuming process	1197.00
p3	RAM use of third most consuming process	1197.00
m3	workload use of third most consuming process	1197.00

TABLE XXII. General description of the laptop system status data used for the c values of the next table. Each event is a measure in a snapshot of system status.

	cpu1	**cpu2**	**cpu3**	**mem**	**p1**	**m1**	**p2**	**m2**	**p3**	**m3**
cpu1	0.00	4.71	4.73	4.45	8.29	11.03	1.81	9.03	3.08	9.16
cpu2	4.71	0.00	2.39	3.16	5.00	9.81	4.47	8.90	3.45	8.54
cpu3	4.73	2.39	0.00	3.15	6.69	13.41	4.23	11.25	4.46	11.88
mem	4.45	3.16	3.15	0.00	3.53	10.29	3.90	8.98	3.14	9.26
p1	8.29	5.00	6.69	3.53	0.00	13.41	8.95	12.61	8.56	12.53
m1	11.03	9.81	13.41	10.29	13.41	0.00	11.63	6.23	9.99	14.90
p2	1.81	4.47	4.23	3.90	8.95	11.63	0.00	10.51	2.82	10.46
m2	9.03	8.90	11.25	8.98	12.61	6.23	10.51	0.00	9.40	1.31
p3	3.08	3.45	4.46	3.14	8.56	9.99	2.82	9.40	0.00	10.30
m3	9.16	8.54	11.88	9.26	12.53	14.90	10.46	1.31	10.30	0.00

TABLE XXIII. Values of c for histograms drawn from laptop system resource status measures.

IV. CONCLUSIONS AND FURTHER BENCHMARKS

The c statistic is robust both to determine if the distributions underlying the samples are the same and to quantify the difference between such probability distributions. The benchmarks for c, given in Section II, are very useful as references to make sense of data e.g. as the example analyses of Section III. Notice that the calculations require no training or clusterization usually involved in classification routines.

The rendering of this article and all tables are automated through Python scripts in order to ease the scrutinization of different settings[3]. After some exploration of the results, this setting was settled as template for this document: simulations used $N_c = 100$ comparisons per measure with $n = n' = 1000$ elements in each sample; all empirical density functions derived from histograms with $N_b = 30$ equally spaced bins. The main reason for this choice is that the results are consistent and stable, but are still of very modest scales. The use of more bins should enhance the results with respect to generality. On the other hand, samples are often not so big, which favors reporting the tables with a small sample size.

ACKNOWLEDGMENTS

Financial support was obtained from CNPq (140860/2013-4, project 870336/1997-5), United Nations Development Program (contract: 2013/000566; project BRA/12/018) and FAPESP. We are also grateful to developers and users of Python scientific tools.

[1] R. Chicheportiche and J.-P. Bouchaud, "Weighted kolmogorov-smirnov test: Accounting for the tails," Physical Review E **86**, 041115 (2012).

[2] G. Deleuze, *Difference and repetition* (Columbia University Press, 1994).

[3] R. Fabbri, "Gmane python package for analysing public email lists," https://pypi.python.org/pypi/gmane (2015).