A distance metric between histograms derived from 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: 5 November 2015)

This document presents reference values for a distance metric derived from the Kolmogorov-Smirnov test statistic $D_{F,F'}$. Each measure of the $D_{F,F'}$ is a distance between two histograms. This distance is normalized by the number of observations in each sample to yield the $c = D_{F,F'}\sqrt{\frac{nn'}{n+n'}}$ 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

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:

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

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

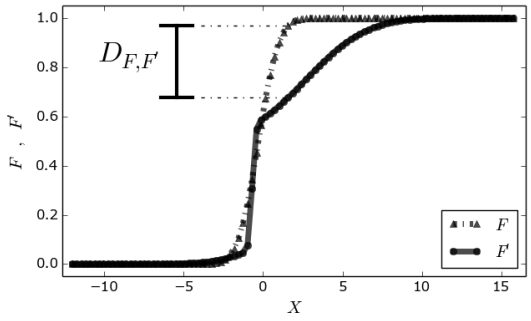

FIG. 1. The Kolmogorov-Smirnov statistic $D_{F,F'}$: the maximum difference between two cumulative distribution functions.

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

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 also 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 $N_c = 100$, each with the sample sizes of $n = 1000$ and $n' = 1000$. 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	8	7	5
5.0	**0.050**	1.36	1	2	2
2.5	**0.025**	1.48	1	0	1
1.0	**0.010**	1.63	1	0	0
0.5	**0.005**	1.73	1	0	0
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, $\mu(c_1) = 0.7775$ and $\sigma(c_1) = 0.2573$, $\mu(c_2) = 0.7976$ and $\sigma(c_2) = 0.2388$, $\mu(c_3) = 0.7596$ and $\sigma(c_3) = 0.2292$.

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	9	4
5.0	**0.050**	1.36	5	4	0
2.5	**0.025**	1.48	2	1	0
1.0	**0.010**	1.63	0	0	0
0.5	**0.005**	1.73	0	0	0
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, $\mu(c_1) = 0.8674$ and $\sigma(c_1) = 0.2468$, $\mu(c_2) = 0.8298$ and $\sigma(c_2) = 0.2608$, $\mu(c_3) = 0.7983$ and $\sigma(c_3) = 0.2205$.

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	9	5	9
5.0	**0.050**	1.36	0	1	3	3
2.5	**0.025**	1.48	0	0	1	1
1.0	**0.010**	1.63	0	0	1	0
0.5	**0.005**	1.73	0	0	1	0
0.1	**0.001**	1.95	0	0	0	0

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, $\mu(c_1) = 0.1107$ and $\sigma(c_1) = 0.0652$, $\mu(c_2) = 0.8079$ and $\sigma(c_2) = 0.2417$, $\mu(c_3) = 0.7775$ and $\sigma(c_3) = 0.2404$. $\mu(c_4) = 0.8209$ and $\sigma(c_4) = 0.2389$.

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	13	10	10	10	9
5.0	**0.050**	1.36	8	3	7	5	7
2.5	**0.025**	1.48	5	2	3	2	2
1.0	**0.010**	1.63	3	0	1	1	1
0.5	**0.005**	1.73	1	0	0	1	1
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, $\mu(c_1) = 0.8379$ and $\sigma(c_1) = 0.3099$, $\mu(c_2) = 0.8392$ and $\sigma(c_2) = 0.2557$, $\mu(c_3) = 0.8642$ and $\sigma(c_3) = 0.2747$. $\mu(c_4) = 0.8383$ and $\sigma(c_4) = 0.2724$. $\mu(c_5) = 0.7920$ and $\sigma(c_5) = 0.2779$.

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. Let D be the KS statistic when sample size goes to infinity. Notably, high values of c yield smaller $|D - D_{F,F'}|$, i.e. more accurate estimates of D through $D_{F,F'}$.

σ	$\mu(c)$	$\sigma(c)$	$min(c)$	$max(c)$	D	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.5	3.978	0.246	3.309,3.354,3.488	4.427,4.517,4.763	0.161	0.178	0.011	1.000	1.000	1.000	1.000	1.000	1.000
0.6	3.084	0.288	2.549,2.571,2.616	3.690,3.734,3.935	0.121	0.138	0.013	1.000	1.000	1.000	1.000	1.000	1.000
0.7	2.296	0.222	1.856,1.878,1.901	2.750,2.750,2.840	0.085	0.103	0.010	1.000	1.000	1.000	1.000	1.000	0.950
0.8	1.614	0.278	0.984,1.140,1.140	2.370,2.370,2.437	0.054	0.072	0.012	0.940	0.870	0.660	0.440	0.300	0.100
0.9	1.123	0.273	0.492,0.559,0.626	1.610,1.744,1.901	0.025	0.050	0.012	0.340	0.230	0.120	0.020	0.020	0.000
1.0	**0.851**	**0.284**	**0.425,0.447,0.447**	**1.521,1.588,1.856**	**0.000**	**0.038**	**0.013**	**0.110**	**0.060**	**0.030**	**0.010**	**0.010**	**0.000**
1.1	1.064	0.258	0.604,0.671,0.671	1.632,1.655,1.923	0.023	0.048	0.012	0.260	0.160	0.060	0.030	0.010	0.000
1.2	1.447	0.279	0.693,0.872,0.939	1.990,2.057,2.057	0.044	0.065	0.012	0.790	0.650	0.400	0.270	0.170	0.040
1.3	1.825	0.262	1.230,1.230,1.275	2.348,2.415,2.750	0.063	0.082	0.012	1.000	0.970	0.900	0.820	0.650	0.290
1.4	2.209	0.292	1.699,1.722,1.766	2.840,2.862,2.862	0.081	0.099	0.013	1.000	1.000	1.000	1.000	0.980	0.790
1.5	2.564	0.301	1.901,1.990,2.012	3.153,3.220,3.309	0.097	0.115	0.013	1.000	1.000	1.000	1.000	1.000	0.990
1.6	2.895	0.279	2.214,2.326,2.348	3.444,3.488,3.578	0.112	0.129	0.012	1.000	1.000	1.000	1.000	1.000	1.000
1.7	3.152	0.290	2.482,2.616,2.616	3.868,3.868,3.891	0.125	0.141	0.013	1.000	1.000	1.000	1.000	1.000	1.000
1.8	3.431	0.266	2.750,2.795,2.974	4.003,4.003,4.025	0.138	0.153	0.012	1.000	1.000	1.000	1.000	1.000	1.000
1.9	3.742	0.293	3.108,3.153,3.198	4.316,4.494,4.763	0.150	0.167	0.013	1.000	1.000	1.000	1.000	1.000	1.000
2.0	3.971	0.269	3.354,3.466,3.511	4.472,4.494,4.584	0.161	0.178	0.012	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	$\mu(c)$	$\sigma(c)$	$min(c)$	$max(c)$	D	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.0	**0.829**	**0.271**	**0.402,0.402,0.402**	**1.565,1.677,1.677**	**0.000**	**0.037**	**0.012**	**0.080**	**0.050**	**0.040**	**0.020**	**0.000**	**0.000**
0.1	1.357	0.414	0.537,0.559,0.581	2.102,2.303,2.415	0.040	0.061	0.019	0.540	0.490	0.340	0.260	0.210	0.110
0.2	2.098	0.451	0.939,1.096,1.118	3.130,3.153,3.198	0.080	0.094	0.020	0.970	0.950	0.910	0.870	0.790	0.640
0.3	3.083	0.399	2.124,2.147,2.191	3.846,3.868,3.935	0.119	0.138	0.018	1.000	1.000	1.000	1.000	1.000	1.000
0.4	3.820	0.477	2.348,2.996,3.130	4.785,4.942,5.411	0.159	0.171	0.021	1.000	1.000	1.000	1.000	1.000	1.000
0.5	4.637	0.449	3.645,3.757,3.801	5.545,5.613,5.680	0.197	0.207	0.020	1.000	1.000	1.000	1.000	1.000	1.000
0.6	5.548	0.483	4.383,4.472,4.584	6.373,6.507,6.507	0.236	0.248	0.022	1.000	1.000	1.000	1.000	1.000	1.000
0.7	6.362	0.415	5.009,5.344,5.568	7.178,7.200,7.357	0.274	0.285	0.019	1.000	1.000	1.000	1.000	1.000	1.000
0.8	7.158	0.441	6.104,6.149,6.306	8.027,8.072,8.408	0.311	0.320	0.020	1.000	1.000	1.000	1.000	1.000	1.000
0.9	7.888	0.419	6.820,6.909,7.133	8.765,8.788,8.967	0.347	0.353	0.019	1.000	1.000	1.000	1.000	1.000	1.000
1.0	8.811	0.413	7.759,7.782,7.938	9.481,9.615,9.839	0.383	0.394	0.018	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	$\mu(c)$	$\sigma(c)$	min(c)	max(c)	D	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.7	6.750	0.377	5.858,5.970,6.037	7.536,7.580,7.647	0.300	0.302	0.017	1.000	1.000	1.000	1.000	1.000	1.000
0.75	5.645	0.303	4.964,5.009,5.076	6.239,6.373,6.619	0.250	0.252	0.014	1.000	1.000	1.000	1.000	1.000	1.000
0.8	4.573	0.287	3.824,3.958,4.047	5.121,5.165,5.299	0.200	0.205	0.013	1.000	1.000	1.000	1.000	1.000	1.000
0.85	3.457	0.234	2.840,2.907,3.019	3.846,4.092,4.114	0.150	0.155	0.010	1.000	1.000	1.000	1.000	1.000	1.000
0.9	2.362	0.251	1.811,1.834,1.901	2.885,2.885,3.019	0.100	0.106	0.011	1.000	1.000	1.000	1.000	1.000	0.930
0.95	1.352	0.255	0.939,0.962,0.962	2.080,2.147,2.214	0.050	0.060	0.011	0.690	0.400	0.270	0.120	0.080	0.030
1.0	**0.797**	**0.235**	**0.358,0.425,0.447**	**1.431,1.453,1.476**	**0.000**	**0.036**	**0.011**	**0.060**	**0.030**	**0.000**	**0.000**	**0.000**	**0.000**
1.05	1.343	0.310	0.805,0.894,0.917	2.147,2.281,2.303	0.048	0.060	0.014	0.550	0.390	0.250	0.180	0.140	0.050
1.1	2.191	0.295	1.565,1.588,1.655	2.706,2.706,2.885	0.091	0.098	0.013	1.000	1.000	1.000	0.980	0.940	0.760
1.15	3.032	0.294	2.370,2.393,2.504	3.645,3.779,3.935	0.130	0.136	0.013	1.000	1.000	1.000	1.000	1.000	1.000
1.2	3.822	0.279	2.885,3.063,3.242	4.360,4.383,4.562	0.167	0.171	0.012	1.000	1.000	1.000	1.000	1.000	1.000
1.25	4.532	0.294	3.801,3.824,3.958	5.188,5.188,5.456	0.200	0.203	0.013	1.000	1.000	1.000	1.000	1.000	1.000
1.3	5.258	0.310	4.383,4.427,4.539	5.903,5.993,6.104	0.231	0.235	0.014	1.000	1.000	1.000	1.000	1.000	1.000
1.35	5.799	0.287	5.098,5.188,5.188	6.261,6.261,6.283	0.259	0.259	0.013	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	$\mu(c)$	$\sigma(c)$	min(c)	max(c)	D	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.5	**0.793**	**0.253**	**0.402,0.425,0.425**	**1.431,1.521,1.610**	**0.000**	**0.035**	**0.011**	**0.080**	**0.030**	**0.020**	**0.000**	**0.000**	**0.000**
0.55	1.701	0.320	1.096,1.118,1.207	2.437,2.549,2.639	0.050	0.076	0.014	0.970	0.850	0.710	0.590	0.450	0.200
0.6	2.813	0.320	2.035,2.191,2.191	3.511,3.600,3.757	0.100	0.126	0.014	1.000	1.000	1.000	1.000	1.000	1.000
0.65	3.972	0.347	3.309,3.309,3.354	4.606,4.629,4.808	0.150	0.178	0.016	1.000	1.000	1.000	1.000	1.000	1.000
0.7	5.073	0.343	4.316,4.405,4.427	5.858,5.858,5.948	0.200	0.227	0.015	1.000	1.000	1.000	1.000	1.000	1.000
0.75	6.205	0.369	5.389,5.613,5.635	7.066,7.290,7.491	0.250	0.278	0.017	1.000	1.000	1.000	1.000	1.000	1.000
0.8	7.238	0.354	6.596,6.641,6.663	8.117,8.117,8.184	0.300	0.324	0.016	1.000	1.000	1.000	1.000	1.000	1.000
0.85	8.337	0.357	7.379,7.647,7.670	9.123,9.190,9.436	0.350	0.373	0.016	1.000	1.000	1.000	1.000	1.000	1.000
0.9	9.425	0.296	8.743,8.855,8.855	10.018,10.085,10.331	0.400	0.422	0.013	1.000	1.000	1.000	1.000	1.000	1.000
0.95	10.520	0.355	9.749,9.883,9.906	11.136,11.158,11.471	0.450	0.470	0.016	1.000	1.000	1.000	1.000	1.000	1.000
1.0	11.661	0.355	10.867,10.957,10.979	12.567,12.611,12.679	0.500	0.521	0.016	1.000	1.000	1.000	1.000	1.000	1.000
1.05	12.722	0.317	11.784,11.918,12.008	13.327,13.349,13.483	0.550	0.569	0.014	1.000	1.000	1.000	1.000	1.000	1.000
1.1	13.856	0.311	13.126,13.215,13.260	14.445,14.467,14.669	0.600	0.620	0.014	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	$\mu(c)$	$\sigma(c)$	min(c)	max(c)	D	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.7	4.604	0.394	3.801,3.868,3.891	5.344,5.501,5.545	0.201	0.206	0.018	1.000	1.000	1.000	1.000	1.000	1.000
0.9	3.310	0.393	2.393,2.460,2.571	4.137,4.181,4.226	0.136	0.148	0.018	1.000	1.000	1.000	1.000	1.000	1.000
1.1	2.233	0.400	1.431,1.498,1.543	3.130,3.153,3.511	0.083	0.100	0.018	1.000	1.000	0.990	0.940	0.880	0.760
1.3	1.336	0.341	0.693,0.738,0.805	1.990,2.147,2.415	0.039	0.060	0.015	0.610	0.410	0.310	0.200	0.150	0.050
1.5	**0.881**	**0.274**	**0.470,0.470,0.492**	**1.632,1.655,1.655**	**0.000**	**0.039**	**0.012**	**0.080**	**0.080**	**0.060**	**0.030**	**0.000**	**0.000**
1.7	1.251	0.319	0.648,0.738,0.760	2.102,2.102,2.124	0.034	0.056	0.014	0.520	0.370	0.210	0.120	0.090	0.030
1.9	1.762	0.344	0.984,0.984,1.185	2.571,2.616,2.862	0.064	0.079	0.015	0.960	0.900	0.810	0.660	0.450	0.260
2.1	2.271	0.345	1.409,1.476,1.543	2.952,3.220,3.421	0.090	0.102	0.015	1.000	1.000	0.980	0.950	0.930	0.840
2.3	2.851	0.378	1.990,2.124,2.191	3.600,3.846,3.913	0.114	0.127	0.017	1.000	1.000	1.000	1.000	1.000	1.000
2.5	3.271	0.371	2.348,2.393,2.594	4.047,4.092,4.114	0.136	0.146	0.017	1.000	1.000	1.000	1.000	1.000	1.000
2.7	3.695	0.367	2.929,3.130,3.130	4.517,4.517,4.808	0.155	0.165	0.016	1.000	1.000	1.000	1.000	1.000	1.000
2.9	4.170	0.379	3.309,3.354,3.399	5.009,5.076,5.098	0.173	0.186	0.017	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	$\mu(c)$	$\sigma(c)$	min(c)	max(c)	D	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
0.7	6.282	0.402	4.919,5.299,5.501	6.909,6.999,7.021	0.274	0.281	0.018	1.000	1.000	1.000	1.000	1.000	1.000
0.9	4.445	0.452	3.511,3.600,3.622	5.344,5.367,5.590	0.186	0.199	0.020	1.000	1.000	1.000	1.000	1.000	1.000
1.1	2.818	0.443	1.588,1.744,1.945	3.734,3.846,4.137	0.114	0.126	0.020	1.000	1.000	1.000	0.990	0.990	0.970
1.3	1.536	0.407	0.783,0.827,0.894	2.415,2.437,2.549	0.053	0.069	0.018	0.750	0.650	0.540	0.350	0.300	0.170
1.5	**0.776**	**0.237**	**0.425,0.447,0.470**	**1.409,1.409,1.565**	**0.000**	**0.035**	**0.011**	**0.060**	**0.040**	**0.010**	**0.000**	**0.000**	**0.000**
1.7	1.499	0.377	0.648,0.738,0.850	2.281,2.326,2.370	0.046	0.067	0.017	0.740	0.660	0.510	0.380	0.300	0.110
1.9	2.246	0.400	0.939,1.386,1.521	2.952,3.063,3.309	0.087	0.100	0.018	0.990	0.990	0.980	0.940	0.870	0.780
2.1	3.051	0.395	2.393,2.393,2.415	3.846,3.891,3.913	0.123	0.136	0.018	1.000	1.000	1.000	1.000	1.000	1.000
2.3	3.710	0.442	2.683,2.795,2.907	4.696,4.718,4.785	0.156	0.166	0.020	1.000	1.000	1.000	1.000	1.000	1.000
2.5	4.415	0.422	3.019,3.175,3.287	5.121,5.188,5.210	0.186	0.197	0.019	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.

C. Changing the sample sizes

Changing the number of elements in each sample changes the value of the c statistic. This section is dedicated to tables in which the c statistic is given for two samples of varied sizes but with fixed underlying distributions. If a same value ϵ is changed of the mean μ or the standard deviation, the first yields greater values of the c statistic. In summary, raising the sample sizes raises c and gives a value of $D_{F,F'}$ closer to D (the maximum difference between the theoretical cumulative distributions).

$n = n'$	$\mu(c)$	$\sigma(c)$	$m(c)$	$min(c)$	$max(c)$	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
100	0.912	0.291	0.919	0.424,0.424,0.424	1.485,1.556,1.768	0.129	0.041	0.140	0.060	0.040	0.010	0.010	0.000
1000	1.283	0.364	1.286	0.626,0.648,0.648	2.080,2.124,2.147	0.057	0.016	0.540	0.390	0.320	0.150	0.100	0.060
10000	3.113	0.462	3.051	2.072,2.128,2.185	4.080,4.278,4.462	0.044	0.007	1.000	1.000	1.000	1.000	1.000	1.000
100000	9.126	0.485	9.076	7.978,8.003,8.128	10.156,10.194,10.252	0.041	0.002	1.000	1.000	1.000	1.000	1.000	1.000

TABLE XI. Measurements of c through simulations with fixed normal distributions but different number of samples. One normal distribution has $\mu = 0$ and $\sigma = 1$. The other normal distribution have $\mu = 0.1$ and $\sigma = 1$. The KS statistic of these distributions converges to 0.0399 as sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	$m(c)$	$min(c)$	$max(c)$	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
100	0.911	0.251	0.849	0.424,0.495,0.495	1.414,1.626,1.697	0.129	0.035	0.130	0.050	0.020	0.010	0.000	0.000
1000	1.466	0.260	1.431	0.917,1.051,1.051	1.990,2.080,2.214	0.066	0.012	0.820	0.620	0.420	0.270	0.180	0.050
10000	3.467	0.243	3.465	2.878,2.970,2.991	3.946,4.080,4.094	0.049	0.003	1.000	1.000	1.000	1.000	1.000	1.000
100000	10.129	0.253	10.125	9.595,9.595,9.631	10.713,10.735,10.896	0.045	0.001	1.000	1.000	1.000	1.000	1.000	1.000

TABLE XII. Measurements of c through simulations with fixed normal distributions but different number of samples. One normal distribution has $\mu = 0$ and $\sigma = 1$. The other normal distribution have $\mu = 0$ and $\sigma = 1.2$. The KS statistic of these distributions converges to 0.0440 as sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
100	0.969	0.295	0.919	0.424,0.495,0.566	1.556,1.626,1.768	0.137	0.042	0.220	0.130	0.100	0.010	0.010	0.000
1000	1.692	0.311	1.677	1.140,1.185,1.207	2.326,2.460,2.661	0.076	0.014	0.970	0.870	0.710	0.560	0.420	0.200
10000	4.156	0.363	4.108	3.345,3.528,3.528	5.077,5.084,5.233	0.059	0.005	1.000	1.000	1.000	1.000	1.000	1.000
100000	11.741	0.255	11.713	11.285,11.301,11.328	12.220,12.274,12.368	0.053	0.001	1.000	1.000	1.000	1.000	1.000	1.000

TABLE XIII. Measurements of c through simulations with fixed uniform distributions but different number of samples. One distribution is uniform in [0,1]. The other distribution is uniform in [0.05,1.05]. The KS statistic of these distributions converges to 0.0500 as sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
100	1.060	0.277	1.061	0.566,0.636,0.636	1.768,1.768,1.838	0.150	0.039	0.260	0.130	0.080	0.040	0.030	0.000
1000	2.050	0.188	2.057	1.677,1.699,1.744	2.482,2.594,2.661	0.092	0.008	1.000	1.000	1.000	1.000	0.980	0.670
10000	6.022	0.141	6.032	5.685,5.720,5.756	6.300,6.350,6.371	0.085	0.002	1.000	1.000	1.000	1.000	1.000	1.000
100000	18.742	0.169	18.737	18.329,18.385,18.418	19.069,19.071,19.103	0.084	0.001	1.000	1.000	1.000	1.000	1.000	1.000

TABLE XIV. Measurements of c through simulations with fixed uniform distributions but different number of samples. One distribution is uniform in [0,1]. The other distribution is uniform in [-0.1,1.1]. The KS statistic of these distributions converges to 0.0833 as sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
100	0.824	0.231	0.778	0.424,0.424,0.495	1.273,1.344,1.414	0.117	0.033	0.060	0.010	0.000	0.000	0.000	0.000
1000	1.211	0.276	1.174	0.760,0.783,0.783	1.878,1.968,1.990	0.054	0.012	0.440	0.280	0.170	0.070	0.050	0.020
10000	2.624	0.358	2.588	1.824,1.945,2.015	3.458,3.642,3.684	0.037	0.005	1.000	1.000	1.000	1.000	1.000	0.980
100000	7.699	0.409	7.744	6.538,6.657,6.878	8.573,8.580,8.765	0.034	0.002	1.000	1.000	1.000	1.000	1.000	1.000

TABLE XV. Measurements of c through simulations with fixed Weibull distributions but different number of samples. One distribution has shape parameter $a = 1.5$. The other distribution has $a = 1.7$. The KS statistic of these distributions converges to 0.0338 as sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	m(c)	min(c)	max(c)	$\mu(D_{F,F'})$	$\sigma(D_{F,F'})$	$\overline{C(0.1)}$	$\overline{C(0.05)}$	$\overline{C(0.025)}$	$\overline{C(0.01)}$	$\overline{C(0.005)}$	$\overline{C(0.001)}$
100	0.853	0.307	0.778	0.354,0.424,0.495	1.556,1.909,1.909	0.121	0.043	0.110	0.080	0.060	0.020	0.020	0.000
1000	1.472	0.385	1.453	0.581,0.604,0.648	2.191,2.303,2.437	0.066	0.017	0.740	0.600	0.470	0.330	0.270	0.130
10000	3.494	0.433	3.500	2.574,2.666,2.673	4.342,4.469,4.518	0.049	0.006	1.000	1.000	1.000	1.000	1.000	1.000
100000	10.506	0.420	10.505	9.539,9.682,9.700	11.259,11.326,11.536	0.047	0.002	1.000	1.000	1.000	1.000	1.000	1.000

TABLE XVI. Measurements of c through simulations with fixed power distributions but different number of samples. One distribution has shape parameter a=1.5. The other distribution has a=1.7. The KS statistic of these distributions converges to 0.0460 as sample sizes increases.

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 word 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. The Bible was the only book where the c statistic is higher between the halves then between the whole text and the halves. This might be due to the differences of the Old and New Testaments.

label	description	chars	tokens	sentences	$\|kw\|$	$\mu(kw)$	$\sigma(kw)$	$\|sw\|$	$\mu(sw)$	$\sigma(sw)$
H,H1,H2	Hamlet by Shakespeare	162881	37360	3106	16722	3.549	1.762	9908	2.721	1.011
B,B1,B2	King James Version of the Holly Bible	4332554	1010654	30103	492901	3.745	1.711	289244	2.927	1.044
M,M1,M2	Moby Dick by Herman Melville	1242990	260819	10059	136008	4.105	2.184	75385	2.847	1.096
E,E1,E2	Esaú e Jacó from Machado de Assis	355706	88472	3822	13984	2.186	1.376	3535	1.486	0.502

TABLE XVII. 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	3.421	3.690	11.560	9.749	12.455	14.847	13.975	13.953	20.321	19.574	19.185
	0.000	0.153	0.165	0.517	0.436	0.557	0.664	0.625	0.624	0.909	0.875	0.858
H1	3.421	0.000	0.693	12.790	11.583	13.260	14.602	13.953	13.864	18.464	17.600	17.307
	0.153	0.000	0.031	0.572	0.518	0.593	0.653	0.624	0.620	0.826	0.787	0.774
H2	3.690	0.693	0.000	13.349	11.717	13.797	15.093	14.378	14.400	18.595	17.712	17.419
	0.165	0.031	0.000	0.597	0.524	0.617	0.675	0.643	0.644	0.832	0.792	0.779
B	11.560	12.790	13.349	0.000	3.980	2.929	15.474	13.685	14.065	22.271	21.888	21.802
	0.517	0.572	0.597	0.000	0.178	0.131	0.692	0.612	0.629	0.996	0.979	0.975
B1	9.749	11.583	11.717	3.980	0.000	5.993	15.451	13.819	14.177	22.159	21.708	21.667
	0.436	0.518	0.524	0.178	0.000	0.268	0.691	0.618	0.634	0.991	0.971	0.969
B2	12.455	13.260	13.797	2.929	5.993	0.000	14.020	12.276	12.321	22.271	21.933	21.846
	0.557	0.593	0.617	0.131	0.268	0.000	0.627	0.549	0.551	0.996	0.981	0.977
M	14.847	14.602	15.093	15.474	15.451	14.020	0.000	1.923	1.789	22.271	21.821	21.779
	0.664	0.653	0.675	0.692	0.691	0.627	0.000	0.086	0.080	0.996	0.976	0.974
M1	13.975	13.953	14.378	13.685	13.819	12.276	1.923	0.000	1.029	22.159	21.686	21.645
	0.625	0.624	0.643	0.612	0.618	0.549	0.086	0.000	0.046	0.991	0.970	0.968
M2	13.953	13.864	14.400	14.065	14.177	12.321	1.789	1.029	0.000	22.181	21.708	21.690
	0.624	0.620	0.644	0.629	0.634	0.551	0.080	0.046	0.000	0.992	0.971	0.970
E	20.321	18.464	18.595	22.271	22.159	22.271	22.271	22.159	22.181	0.000	3.236	2.495
	0.909	0.826	0.832	0.996	0.991	0.996	0.996	0.991	0.992	0.000	0.145	0.112
E1	19.574	17.600	17.712	21.888	21.708	21.933	21.821	21.686	21.708	3.236	0.000	1.395
	0.875	0.787	0.792	0.979	0.971	0.981	0.976	0.970	0.971	0.145	0.000	0.062
E2	19.185	17.307	17.419	21.802	21.667	21.846	21.779	21.645	21.690	2.495	1.395	0.000
	0.858	0.774	0.779	0.975	0.969	0.977	0.974	0.968	0.970	0.112	0.062	0.000

TABLE XVIII. 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	3.868	6.641	5.098	7.491	12.924	11.270	11.203	8.592	11.537	10.554
	0.000	0.126	0.173	0.297	0.228	0.335	0.578	0.504	0.501	0.384	0.516	0.472
H1	2.817	0.000	1.453	8.810	7.312	9.705	13.707	12.634	12.388	6.171	9.050	8.318
	0.126	0.000	0.065	0.394	0.327	0.434	0.613	0.565	0.554	0.276	0.405	0.372
H2	3.868	1.453	0.000	10.308	8.430	10.867	14.646	13.349	13.394	5.292	8.480	7.826
	0.173	0.065	0.000	0.461	0.377	0.486	0.655	0.597	0.599	0.237	0.379	0.350
B	6.641	8.810	10.308	0.000	3.958	2.616	16.122	13.797	14.244	14.641	16.256	15.295
	0.297	0.394	0.461	0.000	0.177	0.117	0.721	0.617	0.637	0.655	0.727	0.684
B1	5.098	7.312	8.430	3.958	0.000	5.970	16.703	14.557	14.825	13.320	15.354	14.378
	0.228	0.327	0.377	0.177	0.000	0.267	0.747	0.651	0.663	0.596	0.687	0.643
B2	7.491	9.705	10.867	2.616	5.970	0.000	14.266	12.254	12.746	15.247	16.529	15.809
	0.335	0.434	0.486	0.117	0.267	0.000	0.638	0.548	0.570	0.682	0.739	0.707
M	12.924	13.707	14.646	16.122	16.703	14.266	0.000	2.326	1.923	17.964	18.512	17.933
	0.578	0.613	0.655	0.721	0.747	0.638	0.000	0.104	0.086	0.803	0.828	0.802
M1	11.270	12.634	13.349	13.797	14.557	12.254	2.326	0.000	0.626	17.155	17.967	17.419
	0.504	0.565	0.597	0.617	0.651	0.548	0.104	0.000	0.028	0.767	0.804	0.779
M2	11.203	12.388	13.394	14.244	14.825	12.746	1.923	0.626	0.000	16.932	17.945	17.285
	0.501	0.554	0.599	0.637	0.663	0.570	0.086	0.028	0.000	0.757	0.803	0.773
E	8.592	6.171	5.292	14.641	13.320	15.247	17.964	17.155	16.932	0.000	4.389	4.031
	0.384	0.276	0.237	0.655	0.596	0.682	0.803	0.767	0.757	0.000	0.196	0.180
E1	11.537	9.050	8.480	16.256	15.354	16.529	18.512	17.967	17.945	4.389	0.000	1.138
	0.516	0.405	0.379	0.727	0.687	0.739	0.828	0.804	0.803	0.196	0.000	0.051
E2	10.554	8.318	7.826	15.295	14.378	15.809	17.933	17.419	17.285	4.031	1.138	0.000
	0.472	0.372	0.350	0.684	0.643	0.707	0.802	0.779	0.773	0.180	0.051	0.000

TABLE XIX. 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
	0.000	0.009	0.009	0.080	0.083	0.078	0.103	0.106	0.101	0.478	0.479	0.478
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
	0.009	0.000	0.017	0.089	0.092	0.086	0.100	0.102	0.098	0.470	0.470	0.469
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
	0.009	0.017	0.000	0.071	0.074	0.069	0.107	0.109	0.104	0.487	0.488	0.487
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
	0.080	0.089	0.071	0.000	0.017	0.016	0.089	0.092	0.087	0.559	0.559	0.558
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
	0.083	0.092	0.074	0.017	0.000	0.033	0.103	0.105	0.101	0.561	0.562	0.561
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
	0.078	0.086	0.069	0.016	0.033	0.000	0.076	0.079	0.074	0.556	0.556	0.556
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
	0.103	0.100	0.107	0.089	0.103	0.076	0.000	0.003	0.003	0.541	0.541	0.540
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
	0.106	0.102	0.109	0.092	0.105	0.079	0.003	0.000	0.006	0.541	0.542	0.541
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
	0.101	0.098	0.104	0.087	0.101	0.074	0.003	0.006	0.000	0.540	0.541	0.540
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
	0.478	0.470	0.487	0.559	0.561	0.556	0.541	0.541	0.540	0.000	0.004	0.004
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
	0.479	0.470	0.488	0.559	0.562	0.556	0.541	0.542	0.541	0.004	0.000	0.007
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
	0.478	0.469	0.487	0.558	0.561	0.556	0.540	0.541	0.540	0.004	0.007	0.000

TABLE XX. 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	3.011	2.550	11.538	10.800	9.816	7.692	6.373	6.082	21.509	21.551	21.525
	0.000	0.135	0.114	0.516	0.483	0.439	0.344	0.285	0.272	0.962	0.964	0.963
H1	3.011	0.000	0.882	11.888	11.567	11.030	9.136	8.219	8.398	19.255	19.297	19.270
	0.135	0.000	0.039	0.532	0.517	0.493	0.409	0.368	0.376	0.861	0.863	0.862
H2	2.550	0.882	0.000	11.261	10.624	10.132	8.343	7.426	7.583	19.700	19.741	19.715
	0.114	0.039	0.000	0.504	0.475	0.453	0.373	0.332	0.339	0.881	0.883	0.882
B	11.538	11.888	11.261	0.000	1.632	1.878	5.724	6.283	7.133	22.338	22.361	22.338
	0.516	0.532	0.504	0.000	0.073	0.084	0.256	0.281	0.319	0.999	1.000	0.999
B1	10.800	11.567	10.624	1.632	0.000	1.140	4.964	5.613	6.507	22.338	22.361	22.334
	0.483	0.517	0.475	0.073	0.000	0.051	0.222	0.251	0.291	0.999	1.000	0.999
B2	9.816	11.030	10.132	1.878	1.140	0.000	4.092	4.629	5.434	22.338	22.361	22.334
	0.439	0.493	0.453	0.084	0.051	0.000	0.183	0.207	0.243	0.999	1.000	0.999
M	7.692	9.136	8.343	5.724	4.964	4.092	0.000	1.722	1.901	22.315	22.338	22.312
	0.344	0.409	0.373	0.256	0.222	0.183	0.000	0.077	0.085	0.998	0.999	0.998
M1	6.373	8.219	7.426	6.283	5.613	4.629	1.722	0.000	1.207	22.292	22.334	22.307
	0.285	0.368	0.332	0.281	0.251	0.207	0.077	0.000	0.054	0.997	0.999	0.998
M2	6.082	8.398	7.583	7.133	6.507	5.434	1.901	1.207	0.000	22.315	22.338	22.312
	0.272	0.376	0.339	0.319	0.291	0.243	0.085	0.054	0.000	0.998	0.999	0.998
E	21.509	19.255	19.700	22.338	22.338	22.338	22.315	22.292	22.315	0.000	2.472	3.751
	0.962	0.861	0.881	0.999	0.999	0.999	0.998	0.997	0.998	0.000	0.111	0.168
E1	21.551	19.297	19.741	22.361	22.361	22.361	22.338	22.334	22.338	2.472	0.000	1.465
	0.964	0.863	0.883	1.000	1.000	1.000	0.999	0.999	0.999	0.111	0.000	0.066
E2	21.525	19.270	19.715	22.338	22.334	22.334	22.312	22.307	22.312	3.751	1.465	0.000
	0.963	0.862	0.882	0.999	0.999	0.999	0.998	0.998	0.998	0.168	0.066	0.000

TABLE XXI. 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	4.327	5.241	7.849	6.015	7.558	8.832	6.909	6.529	20.951	20.970	20.917
	0.000	0.194	0.234	0.351	0.269	0.338	0.395	0.309	0.292	0.937	0.938	0.935
H1	4.327	0.000	1.026	11.322	9.668	11.188	11.859	10.383	10.227	16.623	16.642	16.589
	0.194	0.000	0.046	0.506	0.432	0.500	0.530	0.464	0.457	0.743	0.744	0.742
H2	5.241	1.026	0.000	11.663	10.030	11.596	12.177	10.724	10.567	15.710	15.729	15.676
	0.234	0.046	0.000	0.522	0.449	0.519	0.545	0.480	0.473	0.703	0.703	0.701
B	7.849	11.322	11.663	0.000	2.974	1.476	2.996	2.929	2.862	22.315	22.334	22.281
	0.351	0.506	0.522	0.000	0.133	0.066	0.134	0.131	0.128	0.998	0.999	0.996
B1	6.015	9.668	10.030	2.974	0.000	2.504	4.942	3.130	2.795	22.315	22.334	22.281
	0.269	0.432	0.449	0.133	0.000	0.112	0.221	0.140	0.125	0.998	0.999	0.996
B2	7.558	11.188	11.596	1.476	2.504	0.000	2.639	1.789	1.655	22.338	22.334	22.281
	0.338	0.500	0.519	0.066	0.112	0.000	0.118	0.080	0.074	0.999	0.999	0.996
M	8.832	11.859	12.177	2.996	4.942	2.639	0.000	2.281	2.661	22.338	22.334	22.281
	0.395	0.530	0.545	0.134	0.221	0.118	0.000	0.102	0.119	0.999	0.999	0.996
M1	6.909	10.383	10.724	2.929	3.130	1.789	2.281	0.000	0.738	22.315	22.334	22.281
	0.309	0.464	0.480	0.131	0.140	0.080	0.102	0.000	0.033	0.998	0.999	0.996
M2	6.529	10.227	10.567	2.862	2.795	1.655	2.661	0.738	0.000	22.315	22.334	22.281
	0.292	0.457	0.473	0.128	0.125	0.074	0.119	0.033	0.000	0.998	0.999	0.996
E	20.951	16.623	15.710	22.315	22.315	22.338	22.338	22.315	22.315	0.000	4.870	6.237
	0.937	0.743	0.703	0.998	0.998	0.999	0.999	0.998	0.998	0.000	0.218	0.279
E1	20.970	16.642	15.729	22.334	22.334	22.334	22.334	22.334	22.334	4.870	0.000	1.497
	0.938	0.744	0.703	0.999	0.999	0.999	0.999	0.999	0.999	0.218	0.000	0.067
E2	20.917	16.589	15.676	22.281	22.281	22.281	22.281	22.281	22.281	6.237	1.497	0.000
	0.935	0.742	0.701	0.996	0.996	0.996	0.996	0.996	0.996	0.279	0.067	0.000

TABLE XXII. 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
	0.000	0.014	0.015	0.104	0.114	0.094	0.045	0.047	0.044	0.555	0.556	0.554
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
	0.014	0.000	0.029	0.119	0.128	0.109	0.060	0.061	0.058	0.540	0.541	0.539
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
	0.015	0.029	0.000	0.089	0.099	0.079	0.030	0.032	0.029	0.570	0.570	0.569
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
	0.104	0.119	0.089	0.000	0.012	0.012	0.059	0.057	0.061	0.659	0.660	0.658
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
	0.114	0.128	0.099	0.012	0.000	0.024	0.069	0.067	0.070	0.669	0.669	0.668
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
	0.094	0.109	0.079	0.012	0.024	0.000	0.049	0.048	0.051	0.649	0.650	0.648
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
	0.045	0.060	0.030	0.059	0.069	0.049	0.000	0.003	0.003	0.600	0.601	0.599
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
	0.047	0.061	0.032	0.057	0.067	0.048	0.003	0.000	0.005	0.602	0.602	0.601
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
	0.044	0.058	0.029	0.061	0.070	0.051	0.003	0.005	0.000	0.598	0.599	0.597
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
	0.555	0.540	0.570	0.659	0.669	0.649	0.600	0.602	0.598	0.000	0.011	0.012
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
	0.556	0.541	0.570	0.660	0.669	0.650	0.601	0.602	0.599	0.011	0.000	0.023
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
	0.554	0.539	0.569	0.658	0.668	0.648	0.599	0.601	0.597	0.012	0.023	0.000

TABLE XXIII. 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
W1-1	first wavelet approximation	31
W2-1	higher wavelet leaf	17147
S2	recorded 'front left'	71042
W1-2	first wavelet approximation	32
W2-2	higher wavelet leaf	17771
S3	recorded 'rear center'	65026
W1-3	first wavelet approximation	30
W2-3	higher wavelet leaf	16267
S4	recorded 'rear left'	63010
W1-4	first wavelet approximation	30
W2-4	higher wavelet leaf	15763
S5	noise	67579
W1-5	first wavelet approximation	31
W2-5	higher wavelet leaf	16906

TABLE XXIV. 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	22.671	8.340	2.062	27.502	15.078	2.054	24.858	11.170	2.108	28.275	36.639	1.893	19.829
	0.000	0.321	0.194	0.045	0.365	0.232	0.083	0.375	0.217	0.062	0.385	0.250	0.199	0.340	0.170
W1-1	1.788	0.000	2.371	0.787	0.740	2.970	1.593	0.579	2.657	1.221	2.087	3.085	1.559	1.778	1.362
	0.321	0.000	0.426	0.141	0.186	0.534	0.286	0.148	0.478	0.219	0.534	0.555	0.280	0.452	0.245
W2-1	22.671	2.371	0.000	17.933	2.289	3.937	31.633	2.456	2.242	26.022	3.180	5.087	42.198	2.399	29.940
	0.194	0.426	0.000	0.153	0.405	0.042	0.272	0.449	0.025	0.224	0.581	0.056	0.361	0.431	0.325
S2	8.340	0.787	17.933	0.000	0.596	30.846	23.292	0.791	19.736	16.899	2.369	28.662	41.117	1.819	20.703
	0.045	0.141	0.153	0.000	0.105	0.259	0.126	0.144	0.172	0.092	0.433	0.252	0.221	0.327	0.177
W1-2	2.062	0.740	2.289	0.596	0.000	3.030	1.499	0.730	2.944	0.702	1.664	3.142	1.030	1.540	1.087
	0.365	0.186	0.405	0.105	0.000	0.536	0.265	0.185	0.521	0.124	0.423	0.556	0.182	0.388	0.192
W2-2	27.502	2.970	3.937	30.846	3.030	0.000	36.667	3.046	6.058	30.970	3.404	2.204	47.574	2.512	33.367
	0.232	0.534	0.042	0.259	0.536	0.000	0.310	0.557	0.066	0.263	0.622	0.024	0.401	0.452	0.358
S3	15.078	1.593	31.633	23.292	1.499	36.667	0.000	1.691	32.883	12.918	1.711	36.572	27.398	1.667	24.678
	0.083	0.286	0.272	0.126	0.265	0.310	0.000	0.309	0.288	0.072	0.312	0.325	0.151	0.299	0.213
W1-3	2.054	0.579	2.456	0.791	0.730	3.046	1.691	0.000	3.090	1.236	2.066	3.290	1.330	1.763	1.276
	0.375	0.148	0.449	0.144	0.185	0.557	0.309	0.000	0.565	0.226	0.533	0.601	0.243	0.452	0.233
W2-3	24.858	2.657	2.242	19.736	2.944	6.058	32.883	3.090	0.000	28.223	3.218	4.879	44.388	2.425	31.727
	0.217	0.478	0.025	0.172	0.521	0.066	0.288	0.565	0.000	0.248	0.588	0.055	0.388	0.436	0.348
S4	11.170	1.221	26.022	16.899	0.702	30.970	12.918	1.236	28.223	0.000	1.953	31.182	36.544	1.734	22.988
	0.062	0.219	0.224	0.092	0.124	0.263	0.072	0.226	0.248	0.000	0.357	0.278	0.202	0.311	0.199
W1-4	2.108	2.087	3.180	2.369	1.664	3.404	1.711	2.066	3.218	1.953	0.000	3.431	2.272	1.377	2.746
	0.385	0.534	0.581	0.433	0.423	0.622	0.312	0.533	0.588	0.357	0.000	0.627	0.415	0.353	0.502
W2-4	28.275	3.085	5.087	28.662	3.142	2.204	36.572	3.290	4.879	31.182	3.431	0.000	47.707	2.512	34.399
	0.250	0.555	0.056	0.252	0.556	0.024	0.325	0.601	0.055	0.278	0.627	0.000	0.422	0.452	0.381
S5	36.639	1.559	42.198	41.117	1.030	47.574	27.398	1.330	44.388	36.544	2.272	47.707	0.000	2.322	17.089
	0.199	0.280	0.361	0.221	0.182	0.401	0.151	0.243	0.388	0.202	0.415	0.422	0.000	0.417	0.147
W1-5	1.893	1.778	2.399	1.819	1.540	2.512	1.667	1.763	2.425	1.734	1.377	2.512	2.322	0.000	2.503
	0.340	0.452	0.431	0.327	0.388	0.452	0.299	0.452	0.436	0.311	0.353	0.452	0.417	0.000	0.450
W2-5	19.829	1.362	29.940	20.703	1.087	33.367	24.678	1.276	31.727	22.988	2.746	34.399	17.089	2.503	0.000
	0.170	0.245	0.325	0.177	0.192	0.358	0.213	0.233	0.348	0.199	0.502	0.381	0.147	0.450	0.000

TABLE XXV. 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
Bach1	BWV735 from J. S. Bach	236
Bach2	BWV648 from J. S. Bach	272
Moza1	K80 from W. A. Mozart	538
Moza2	K458 from W. A. Mozart	4218
Beet1	Opus 18, n1, mov. 3 from L. van Beethoven	1289
Beet2	Opus 132 from L. van Beethoven	17884
Schön	Opus 19, mov. 2 from A. Schönberg	102

TABLE XXVI. 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 0.00	1.88 0.14	1.89 0.13	2.60 0.15	4.12 0.17	4.43 0.21	5.49 0.21	2.62 0.28
Bach1	1.88 0.14	0.00 0.00	1.00 0.09	1.27 0.10	1.50 0.10	2.09 0.15	2.51 0.16	1.54 0.18
Bach2	1.89 0.13	1.00 0.09	0.00 0.00	1.26 0.09	1.78 0.11	2.20 0.15	2.73 0.17	1.52 0.18
Moza1	2.60 0.15	1.27 0.10	1.26 0.09	0.00 0.00	2.14 0.10	2.08 0.11	2.25 0.10	1.79 0.19
Moza2	4.12 0.17	1.50 0.10	1.78 0.11	2.14 0.10	0.00 0.00	2.99 0.10	5.52 0.09	2.02 0.20
Beet1	4.43 0.21	2.09 0.15	2.20 0.15	2.08 0.11	2.99 0.10	0.00 0.00	2.34 0.07	2.31 0.24
Beet2	5.49 0.21	2.51 0.16	2.73 0.17	2.25 0.10	5.52 0.09	2.34 0.07	0.00 0.00	2.39 0.24
Schön	2.62 0.28	1.54 0.18	1.52 0.18	1.79 0.19	2.02 0.20	2.31 0.24	2.39 0.24	0.00 0.00

TABLE XXVII. 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
cpu2	workload of the second most active processor	422
cpu3	workload of the third most active processor	1046
mem	RAM use in kB	557
p1	workload use of most consuming process	1197
m1	RAM use of most consuming process	1197
p2	workload use of second most consuming process	1197
m2	RAM use of second most consuming process	1197
p3	RAM use of third most consuming process	1197
m3	workload use of third most consuming process	1197

TABLE XXVIII. 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 0.00	5.13 0.30	4.73 0.22	4.84 0.26	8.57 0.38	11.03 0.49	1.84 0.08	9.25 0.41	3.39 0.15	9.94 0.44
cpu2	5.13 0.30	0.00 0.00	2.97 0.17	3.83 0.25	8.89 0.50	9.90 0.56	4.72 0.27	9.03 0.51	3.64 0.21	9.00 0.51
cpu3	4.73 0.22	2.97 0.17	0.00 0.00	3.69 0.19	8.63 0.37	13.44 0.57	4.63 0.20	12.50 0.53	4.51 0.19	12.00 0.51
mem	4.84 0.26	3.83 0.25	3.69 0.19	0.00 0.00	4.16 0.21	10.77 0.55	4.60 0.24	9.53 0.49	3.97 0.20	9.48 0.49
p1	8.57 0.38	8.89 0.50	8.63 0.37	4.16 0.21	0.00 0.00	13.75 0.56	10.08 0.41	12.77 0.52	9.85 0.40	12.53 0.51
m1	11.03 0.49	9.90 0.56	13.44 0.57	10.77 0.55	13.75 0.56	0.00 0.00	12.26 0.50	12.53 0.51	10.57 0.43	15.43 0.63
p2	1.84 0.08	4.72 0.27	4.63 0.20	4.60 0.24	10.08 0.41	12.26 0.50	0.00 0.00	10.51 0.43	2.82 0.12	10.85 0.44
m2	9.25 0.41	9.03 0.51	12.50 0.53	9.53 0.49	12.77 0.52	12.53 0.51	10.51 0.43	0.00 0.00	10.30 0.42	8.67 0.35
p3	3.39 0.15	3.64 0.21	4.51 0.19	3.97 0.20	9.85 0.40	10.57 0.43	2.82 0.12	10.30 0.42	0.00 0.00	10.30 0.42
m3	9.94 0.44	9.00 0.51	12.00 0.51	9.48 0.49	12.53 0.51	15.43 0.63	10.85 0.44	8.67 0.35	10.30 0.42	0.00 0.00

TABLE XXIX. 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).