

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

Renato Fabbri<sup>1, a)</sup>

São Carlos Institute of Physics, University of São Paulo (IFSC/USP), PO Box 369, 13560-970, São Carlos, SP, Brazil

(Dated: 4 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  $\alpha$  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

<b>I. Introduction</b>	1
A. Philosophical and technological note	2
B. Document outline	2
<b>II. References through simulations</b>	2
A. When the null hypothesis is true	2
B. When the null hypothesis is false	4
C. Changing the sample sizes	7
<b>III. Example uses in empirical data</b>	9
A. Text	9
B. Audio	13
C. Music	15
D. OS status	15
<b>IV. Conclusions and further benchmarks</b>	15
<b>Acknowledgments</b>	16

## 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:

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

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

$\alpha$	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:

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

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

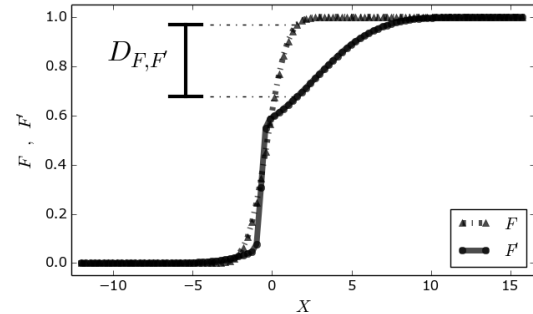


FIG. 1. The Kolmogorov-Smirnov statistic  $D_{F,F'}$ : the maximum difference between two cumulative distribution functions.

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<sup>1</sup>. 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  $c$ .

<sup>a)</sup> <http://ifsc.usp.br/~fabbri/>; Electronic mail: [fabbri@usp.br](mailto: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<sup>2</sup>. 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  $\alpha N_c$ . 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  $\alpha$ . This section reports that  $|C(\alpha)|$  very rarely exceeds  $\alpha N_c$ , for all probability distributions and settings. Also important are that  $c > c(\alpha)$  in many cases and that the tabulated  $\alpha$  values are also good estimates of the upper limit of the frequency of such an event.

$\alpha N_c$	$\alpha$	$c(\alpha)$	$ C_1(\alpha) $	$ C_2(\alpha) $	$ C_3(\alpha) $
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  $\alpha N_c$  of rejections of the null hypothesis for significance levels  $\alpha$ . 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  $\sigma(c_3) = 0.2292$ .

$\alpha N_c$	$\alpha$	$c(\alpha)$	$ C_1(\alpha) $	$ C_2(\alpha) $	$ C_3(\alpha) $
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  $\alpha N_c$  of rejections of the null hypothesis for critical values of  $\alpha$ . 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  $\sigma(c_3) = 0.2205$ .

$\alpha N_c$	$\alpha$	$c(\alpha)$	$ C_1(\alpha) $	$ C_2(\alpha) $	$ C_3(\alpha) $	$ C_4(\alpha) $
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  $\alpha N_c$  of rejections of the null hypothesis for critical values of  $\alpha$ . 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  $\sigma(c_4) = 0.2389$ .

$\alpha N_c$	$\alpha$	$c(\alpha)$	$ C_1(\alpha) $	$ C_2(\alpha) $	$ C_3(\alpha) $	$ C_4(\alpha) $	$ C_5(\alpha) $
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  $\alpha N_c$  of rejections of the null hypothesis for critical values of  $\alpha$ . 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 is false**

The null hypothesis is always false for a sufficiently small significance level  $\alpha$ . 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  $\alpha$ . The null hypothesis is true in the boldface lines.

$\sigma$	$\mu(c)$	$\sigma(c)$	$\min(c)$	$\max(c)$	$D$	$\mu(D_{n,n'})$	$\sigma(D_{n,n'})$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
0.5	4.002	0.295	3.130,3.399,3.466	4.718,4.785,4.852	0.161	0.179	0.013	1.000	1.000	1.000	1.000	1.000	1.000
0.6	3.122	0.273	2.415,2.504,2.616	3.645,3.690,4.159	0.121	0.140	0.012	1.000	1.000	1.000	1.000	1.000	1.000
0.7	2.303	0.319	1.565,1.677,1.722	2.996,3.041,3.175	0.085	0.103	0.014	1.000	1.000	1.000	0.990	0.970	0.860
0.8	1.651	0.268	1.118,1.140,1.163	2.214,2.214,2.281	0.054	0.074	0.012	0.950	0.880	0.700	0.530	0.390	0.150
0.9	1.093	0.291	0.470,0.559,0.604	1.699,1.856,2.080	0.025	0.049	0.013	0.260	0.160	0.100	0.050	0.020	0.010
<b>1.0</b>	<b>0.799</b>	<b>0.253</b>	<b>0.335,0.402,0.425</b>	<b>1.386,1.498,1.588</b>	<b>0.000</b>	<b>0.036</b>	<b>0.011</b>	<b>0.070</b>	<b>0.030</b>	<b>0.020</b>	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>
1.1	1.011	0.248	0.514,0.559,0.604	1.588,1.588,1.968	0.023	0.045	0.011	0.190	0.080	0.040	0.010	0.010	0.010
1.2	1.454	0.261	0.783,0.962,0.984	1.990,2.012,2.393	0.044	0.065	0.012	0.800	0.670	0.450	0.250	0.120	0.040
1.3	1.843	0.286	1.140,1.140,1.275	2.415,2.460,2.527	0.063	0.082	0.013	0.980	0.970	0.900	0.760	0.610	0.340
1.4	2.189	0.269	1.610,1.677,1.699	2.929,2.929,3.019	0.081	0.098	0.012	1.000	1.000	1.000	0.990	0.950	0.840
1.5	2.593	0.290	1.811,1.856,1.856	3.153,3.242,3.265	0.097	0.116	0.013	1.000	1.000	1.000	1.000	1.000	0.970
1.6	2.847	0.265	2.281,2.393,2.415	3.488,3.511,3.533	0.112	0.127	0.012	1.000	1.000	1.000	1.000	1.000	1.000
1.7	3.241	0.267	2.639,2.706,2.817	3.824,3.891,4.293	0.125	0.145	0.012	1.000	1.000	1.000	1.000	1.000	1.000
1.8	3.478	0.274	2.929,3.019,3.041	4.092,4.181,4.226	0.138	0.156	0.012	1.000	1.000	1.000	1.000	1.000	1.000
1.9	3.737	0.284	2.996,2.996,3.198	4.293,4.338,4.360	0.150	0.167	0.013	1.000	1.000	1.000	1.000	1.000	1.000
2.0	3.960	0.293	3.309,3.354,3.466	4.562,4.785,5.054	0.161	0.177	0.013	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 against normal distributions with  $\mu = 0$  and different values of  $\sigma$ .

$\mu$	$\mu(c)$	$\sigma(c)$	$\min(c)$	$\max(c)$	$D$	$\mu(D_{n,n'})$	$\sigma(D_{n,n'})$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
<b>0.0</b>	<b>0.872</b>	<b>0.249</b>	<b>0.447,0.447,0.470</b>	<b>1.386,1.521,1.565</b>	<b>0.000</b>	<b>0.039</b>	<b>0.011</b>	<b>0.100</b>	<b>0.050</b>	<b>0.020</b>	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>
0.1	1.266	0.376	0.514,0.559,0.626	2.124,2.147,2.236	0.040	0.057	0.017	0.540	0.390	0.290	0.190	0.110	0.050
0.2	2.211	0.472	0.805,1.163,1.319	3.265,3.309,3.376	0.080	0.099	0.021	0.980	0.950	0.930	0.890	0.840	0.720
0.3	3.033	0.475	1.699,1.766,1.834	3.935,4.003,4.137	0.119	0.136	0.021	1.000	1.000	1.000	1.000	0.990	0.970
0.4	3.770	0.466	2.393,2.750,2.795	4.584,4.606,4.696	0.159	0.169	0.021	1.000	1.000	1.000	1.000	1.000	1.000
0.5	4.656	0.440	3.645,3.734,3.846	5.501,5.568,5.657	0.197	0.208	0.020	1.000	1.000	1.000	1.000	1.000	1.000
0.6	5.539	0.420	4.584,4.740,4.763	6.507,6.529,6.708	0.236	0.248	0.019	1.000	1.000	1.000	1.000	1.000	1.000
0.7	6.374	0.436	5.054,5.523,5.590	7.178,7.334,7.424	0.274	0.285	0.019	1.000	1.000	1.000	1.000	1.000	1.000
0.8	7.169	0.383	6.306,6.395,6.440	8.050,8.273,8.363	0.311	0.321	0.017	1.000	1.000	1.000	1.000	1.000	1.000
0.9	7.968	0.436	6.708,6.887,7.066	8.698,8.810,8.922	0.347	0.356	0.020	1.000	1.000	1.000	1.000	1.000	1.000
1.0	8.780	0.404	7.916,7.960,8.027	9.772,9.883,9.906	0.383	0.393	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 against normal distributions with different values of  $\mu$  and fixed  $\sigma = 1$ .

$b$	$\mu(c)$	$\sigma(c)$	$\min(c)$	$\max(c)$	$D$	$\mu(D_{n,n'})$	$\sigma(D_{n,n'})$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
0.7	6.820	0.320	5.970,6.261,6.306	7.491,7.580,7.603	0.300	0.305	0.014	1.000	1.000	1.000	1.000	1.000	1.000
0.75	5.611	0.336	4.629,4.830,4.897	6.239,6.440,6.462	0.250	0.251	0.015	1.000	1.000	1.000	1.000	1.000	1.000
0.8	4.531	0.320	3.935,3.935,3.980	5.232,5.277,5.322	0.200	0.203	0.014	1.000	1.000	1.000	1.000	1.000	1.000
0.85	3.424	0.282	2.750,2.750,2.817	3.913,4.003,4.003	0.150	0.153	0.013	1.000	1.000	1.000	1.000	1.000	1.000
0.9	2.369	0.254	1.878,1.901,1.901	2.862,3.019,3.242	0.100	0.106	0.011	1.000	1.000	1.000	1.000	1.000	0.940
0.95	1.402	0.244	0.872,0.917,1.051	1.945,2.035,2.303	0.050	0.063	0.011	0.800	0.540	0.290	0.190	0.080	0.020
<b>1.0</b>	<b>0.838</b>	<b>0.281</b>	<b>0.335,0.425,0.492</b>	<b>1.632,1.699,1.699</b>	<b>0.000</b>	<b>0.037</b>	<b>0.013</b>	<b>0.110</b>	<b>0.090</b>	<b>0.030</b>	<b>0.030</b>	<b>0.000</b>	<b>0.000</b>
1.05	1.406	0.287	0.760,0.827,0.939	2.035,2.080,2.281	0.048	0.063	0.013	0.690	0.540	0.370	0.230	0.130	0.030
1.1	2.181	0.268	1.632,1.677,1.699	2.795,2.840,2.996	0.091	0.098	0.012	1.000	1.000	1.000	1.000	0.960	0.790
1.15	3.038	0.254	2.415,2.482,2.527	3.511,3.600,3.622	0.130	0.136	0.011	1.000	1.000	1.000	1.000	1.000	1.000
1.2	3.796	0.277	3.220,3.242,3.265	4.360,4.383,4.405	0.167	0.170	0.012	1.000	1.000	1.000	1.000	1.000	1.000
1.25	4.514	0.282	3.824,3.868,3.913	5.076,5.121,5.277	0.200	0.202	0.013	1.000	1.000	1.000	1.000	1.000	1.000
1.3	5.195	0.278	4.494,4.517,4.539	5.657,5.747,5.814	0.231	0.232	0.012	1.000	1.000	1.000	1.000	1.000	1.000
1.35	5.788	0.356	4.718,4.763,4.830	6.373,6.418,6.708	0.259	0.259	0.016	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.

$\mu$	$\mu(c)$	$\sigma(c)$	min(c)	max(c)	$D$	$\mu(D_{n,n'})$	$\sigma(D_{n,n'})$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
<b>0.5</b>	<b>0.837</b>	<b>0.277</b>	<b>0.380,0.425,0.447</b>	<b>1.453,1.565,1.834</b>	<b>0.000</b>	<b>0.037</b>	<b>0.012</b>	<b>0.120</b>	<b>0.060</b>	<b>0.020</b>	<b>0.010</b>	<b>0.010</b>	<b>0.000</b>
0.55	1.697	0.317	0.984,1.096,1.140	2.415,2.482,2.750	0.050	0.076	0.014	0.970	0.850	0.770	0.600	0.400	0.170
0.6	2.844	0.338	2.147,2.258,2.258	3.555,3.622,3.935	0.100	0.127	0.015	1.000	1.000	1.000	1.000	1.000	1.000
0.65	3.954	0.342	3.287,3.444,3.466	4.763,5.143,5.322	0.150	0.177	0.015	1.000	1.000	1.000	1.000	1.000	1.000
0.7	4.978	0.347	4.204,4.316,4.338	5.836,5.926,5.970	0.200	0.223	0.016	1.000	1.000	1.000	1.000	1.000	1.000
0.75	6.047	0.333	5.299,5.344,5.389	6.753,6.775,6.865	0.250	0.270	0.015	1.000	1.000	1.000	1.000	1.000	1.000
0.8	7.215	0.364	6.306,6.529,6.552	8.027,8.050,8.095	0.300	0.323	0.016	1.000	1.000	1.000	1.000	1.000	1.000
0.85	8.383	0.307	7.849,7.871,7.871	9.078,9.078,9.190	0.350	0.375	0.014	1.000	1.000	1.000	1.000	1.000	1.000
0.9	9.450	0.359	8.408,8.721,8.721	10.152,10.196,10.196	0.400	0.423	0.016	1.000	1.000	1.000	1.000	1.000	1.000
0.95	10.524	0.419	9.749,9.772,9.839	11.471,11.516,11.605	0.450	0.471	0.019	1.000	1.000	1.000	1.000	1.000	1.000
1.0	11.640	0.312	10.912,10.979,11.024	12.365,12.544,12.611	0.500	0.521	0.014	1.000	1.000	1.000	1.000	1.000	1.000
1.05	12.721	0.367	11.874,11.941,11.963	13.349,13.349,13.506	0.550	0.569	0.016	1.000	1.000	1.000	1.000	1.000	1.000
1.1	13.795	0.273	13.103,13.238,13.372	14.490,14.490,14.602	0.600	0.617	0.012	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)	$D$	$\mu(D_{n,n'})$	$\sigma(D_{n,n'})$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
0.01	0.031	0.013	0.022,0.022,0.022	0.067,0.067,0.089	nan	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
0.1	0.364	0.222	0.045,0.067,0.067	0.783,0.984,1.431	nan	0.016	0.010	0.010	0.010	0.000	0.000	0.000	0.000
0.3	5.064	0.713	1.789,2.996,3.019	6.127,6.127,6.261	nan	0.226	0.032	1.000	1.000	1.000	1.000	1.000	0.990
0.5	6.114	0.382	5.232,5.299,5.434	6.909,7.088,7.133	nan	0.273	0.017	1.000	1.000	1.000	1.000	1.000	1.000
0.7	4.498	0.367	3.622,3.645,3.779	5.143,5.255,5.434	nan	0.201	0.016	1.000	1.000	1.000	1.000	1.000	1.000
0.9	3.184	0.364	2.460,2.482,2.482	3.891,3.958,3.980	nan	0.142	0.016	1.000	1.000	1.000	1.000	1.000	1.000
1.1	2.115	0.323	1.409,1.565,1.588	2.750,2.817,2.952	nan	0.095	0.014	1.000	1.000	0.990	0.930	0.850	0.670
1.3	1.315	0.285	0.648,0.783,0.850	1.945,2.012,2.102	nan	0.059	0.013	0.590	0.420	0.260	0.170	0.090	0.020
<b>1.5</b>	<b>0.847</b>	<b>0.275</b>	<b>0.402,0.470,0.470</b>	<b>1.543,1.610,1.722</b>	<b>nan</b>	<b>0.038</b>	<b>0.012</b>	<b>0.110</b>	<b>0.050</b>	<b>0.030</b>	<b>0.010</b>	<b>0.000</b>	<b>0.000</b>
1.7	1.195	0.384	0.581,0.604,0.648	2.281,2.326,2.549	nan	0.053	0.017	0.370	0.300	0.210	0.150	0.090	0.030
1.9	1.681	0.361	0.939,0.984,1.029	2.549,2.639,2.885	nan	0.075	0.016	0.910	0.820	0.720	0.580	0.420	0.150
2.1	2.239	0.383	1.364,1.409,1.431	2.996,3.019,3.130	nan	0.100	0.017	1.000	1.000	0.970	0.960	0.910	0.730
2.3	2.739	0.356	1.968,2.102,2.124	3.488,3.622,3.712	nan	0.122	0.016	1.000	1.000	1.000	1.000	1.000	1.000
2.5	3.308	0.352	2.594,2.639,2.773	3.980,4.025,4.405	nan	0.148	0.016	1.000	1.000	1.000	1.000	1.000	1.000
2.7	3.666	0.424	2.728,2.862,2.952	4.606,4.763,4.875	nan	0.164	0.019	1.000	1.000	1.000	1.000	1.000	1.000
2.9	4.043	0.362	3.063,3.265,3.332	4.740,4.785,4.964	nan	0.181	0.016	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)	$D$	$\mu(D_{n,n'})$	$\sigma(D_{n,n'})$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
0.7	6.368	0.378	5.478,5.545,5.635	7.245,7.401,7.468	0.274	0.285	0.017	1.000	1.000	1.000	1.000	1.000	1.000
0.9	4.374	0.453	3.019,3.332,3.511	5.232,5.255,5.344	0.186	0.196	0.020	1.000	1.000	1.000	1.000	1.000	1.000
1.1	2.863	0.405	1.789,2.012,2.080	3.801,3.891,4.070	0.114	0.128	0.018	1.000	1.000	1.000	1.000	1.000	0.990
1.3	1.610	0.456	0.537,0.604,0.671	2.437,2.527,2.683	0.053	0.072	0.020	0.810	0.740	0.630	0.510	0.430	0.260
<b>1.5</b>	<b>0.818</b>	<b>0.225</b>	<b>0.402,0.447,0.470</b>	<b>1.386,1.386,1.409</b>	<b>0.000</b>	<b>0.037</b>	<b>0.010</b>	<b>0.050</b>	<b>0.030</b>	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>
1.7	1.378	0.355	0.470,0.537,0.671	2.102,2.147,2.236	0.046	0.062	0.016	0.700	0.540	0.370	0.190	0.160	0.080
1.9	2.263	0.453	1.029,1.319,1.342	3.041,3.175,3.488	0.087	0.101	0.020	0.990	0.970	0.970	0.900	0.840	0.750
2.1	2.959	0.445	1.990,2.080,2.169	3.734,3.779,3.935	0.123	0.132	0.020	1.000	1.000	1.000	1.000	1.000	1.000
2.3	3.811	0.467	2.616,2.728,2.795	4.718,4.740,4.808	0.156	0.170	0.021	1.000	1.000	1.000	1.000	1.000	1.000
2.5	4.411	0.437	3.354,3.421,3.555	5.232,5.389,5.545	0.186	0.197	0.020	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  $\epsilon$  is changed of the mean  $\mu$  or the standard deviation, the first yields greater values of the  $c$  statistic.

$n = n'$	$\mu(c)$	$\sigma(c)$	$m(c)$	$\min(c)$	$\max(c)$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
<b>100</b>	0.809	0.255	0.778	0.354,0.424,0.424	1.414,1.485,1.838	0.060	0.030	0.020	0.010	0.010	0.000
<b>1000</b>	1.326	0.401	1.241	0.581,0.738,0.738	2.124,2.303,2.885	0.530	0.430	0.330	0.180	0.140	0.080
<b>10000</b>	3.137	0.368	3.125	2.326,2.376,2.432	3.833,3.939,3.967	1.000	1.000	1.000	1.000	1.000	1.000
<b>100000</b>	9.118	0.464	9.127	7.960,8.137,8.150	9.866,10.221,10.541	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 0.04 when sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	$m(c)$	$\min(c)$	$\max(c)$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
<b>100</b>	0.874	0.301	0.849	0.424,0.424,0.424	1.626,1.697,1.768	0.150	0.080	0.070	0.020	0.010	0.000
<b>1000</b>	1.514	0.286	1.509	0.939,0.939,0.962	2.080,2.169,2.258	0.830	0.710	0.540	0.360	0.230	0.060
<b>10000</b>	3.476	0.249	3.493	2.984,2.998,3.005	3.910,4.016,4.306	1.000	1.000	1.000	1.000	1.000	1.000
<b>100000</b>	10.139	0.296	10.118	9.405,9.492,9.537	10.679,10.695,10.878	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 0.04 when sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	$m(c)$	$\min(c)$	$\max(c)$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
<b>100</b>	0.981	0.334	0.919	0.424,0.424,0.495	1.697,1.697,2.051	0.200	0.140	0.120	0.050	0.010	0.010
<b>1000</b>	1.710	0.361	1.677	0.962,1.006,1.140	2.482,2.527,2.706	0.930	0.850	0.730	0.540	0.410	0.210
<b>10000</b>	4.119	0.344	4.062	3.571,3.578,3.599	4.886,4.950,5.204	1.000	1.000	1.000	1.000	1.000	1.000
<b>100000</b>	11.744	0.325	11.709	11.093,11.167,11.254	12.524,12.705,13.099	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 0.10 when sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	$m(c)$	$\min(c)$	$\max(c)$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
<b>100</b>	1.013	0.241	0.990	0.495,0.566,0.566	1.485,1.556,1.556	0.220	0.090	0.030	0.000	0.000	0.000
<b>1000</b>	2.076	0.205	2.057	1.699,1.699,1.722	2.571,2.750,2.840	1.000	1.000	1.000	1.000	0.950	0.770
<b>10000</b>	6.026	0.164	6.014	5.657,5.671,5.742	6.392,6.421,6.442	1.000	1.000	1.000	1.000	1.000	1.000
<b>100000</b>	18.740	0.146	18.740	18.383,18.427,18.499	19.060,19.062,19.172	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 0.05 when sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	$m(c)$	$\min(c)$	$\max(c)$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
<b>100</b>	0.856	0.257	0.778	0.424,0.424,0.424	1.414,1.485,1.626	0.100	0.030	0.020	0.000	0.000	0.000
<b>1000</b>	1.183	0.282	1.129	0.604,0.648,0.693	1.878,1.901,1.901	0.430	0.280	0.130	0.070	0.050	0.000
<b>10000</b>	2.557	0.347	2.475	1.895,1.916,1.945	3.373,3.408,3.543	1.000	1.000	1.000	1.000	1.000	0.970
<b>100000</b>	7.673	0.433	7.641	6.701,6.802,6.956	8.428,8.649,8.955	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 0.13 when sample sizes increases.

$n = n'$	$\mu(c)$	$\sigma(c)$	$m(c)$	$\min(c)$	$\max(c)$	$C(0.1)$	$C(0.05)$	$C(0.025)$	$C(0.01)$	$C(0.005)$	$C(0.001)$
<b>100</b>	0.935	0.303	0.919	0.495,0.495,0.495	1.768,1.838,1.909	0.120	0.100	0.080	0.030	0.030	0.000
<b>1000</b>	1.468	0.404	1.465	0.470,0.626,0.716	2.281,2.281,2.571	0.730	0.650	0.480	0.380	0.240	0.120
<b>10000</b>	3.581	0.412	3.557	2.715,2.821,2.885	4.455,4.462,4.540	1.000	1.000	1.000	1.000	1.000	1.000
<b>100000</b>	10.469	0.470	10.489	9.284,9.360,9.472	11.337,11.464,11.623	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 0.05 when 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 than 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 $	$\mu(sw)$	$\sigma(sw)$
<b>H,H1,H2</b>	Hamlet by Shakespeare	162881	37360	3106	16722	3.549	1.762	9908	2.721	1.011
<b>B,B1,B2</b>	King James Version of the Holly Bible	4332554	1010654	30103	492901	3.745	1.711	289244	2.927	1.044
<b>M,M1,M2</b>	Moby Dick by Herman Melville	1242990	260819	10059	136008	4.105	2.184	75385	2.847	1.096
<b>E,E1,E2</b>	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.

	<b>H</b>	<b>H1</b>	<b>H2</b>	<b>B</b>	<b>B1</b>	<b>B2</b>	<b>M</b>	<b>M1</b>	<b>M2</b>	<b>E</b>	<b>E1</b>	<b>E2</b>
<b>H</b>	0.000	3.421	3.690	11.560	9.749	12.455	14.847	13.975	13.953	20.321	19.574	19.185
<b>H1</b>	3.421	0.000	0.693	12.790	11.583	13.260	14.602	13.953	13.864	18.464	17.600	17.307
<b>H2</b>	3.690	0.693	0.000	13.349	11.717	13.797	15.093	14.378	14.400	18.595	17.712	17.419
<b>B</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
<b>B1</b>	9.749	11.583	11.717	3.980	0.000	5.993	15.451	13.819	14.177	22.159	21.708	21.667
<b>B2</b>	12.455	13.260	13.797	2.929	5.993	0.000	14.020	12.276	12.321	22.271	21.933	21.846
<b>M</b>	14.847	14.602	15.093	15.474	15.451	14.020	0.000	1.923	1.789	22.271	21.821	21.779
<b>M1</b>	13.975	13.953	14.378	13.685	13.819	12.276	1.923	0.000	1.029	22.159	21.686	21.645
<b>M2</b>	13.953	13.864	14.400	14.065	14.177	12.321	1.789	1.029	0.000	22.181	21.708	21.690
<b>E</b>	20.321	18.464	18.595	22.271	22.159	22.271	22.271	22.159	22.181	0.000	3.236	2.495
<b>E1</b>	19.574	17.600	17.712	21.888	21.708	21.933	21.821	21.686	21.708	3.236	0.000	1.395
<b>E2</b>	19.185	17.307	17.419	21.802	21.667	21.846	21.779	21.645	21.690	2.495	1.395	0.000

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

	<b>H</b>	<b>H1</b>	<b>H2</b>	<b>B</b>	<b>B1</b>	<b>B2</b>	<b>M</b>	<b>M1</b>	<b>M2</b>	<b>E</b>	<b>E1</b>	<b>E2</b>
<b>H</b>	0.000	2.817	3.868	6.641	5.098	7.491	12.924	11.270	11.203	8.592	11.537	10.554
<b>H1</b>	2.817	0.000	1.453	8.810	7.312	9.705	13.707	12.634	12.388	6.171	9.050	8.318
<b>H2</b>	3.868	1.453	0.000	10.308	8.430	10.867	14.646	13.349	13.394	5.292	8.480	7.826
<b>B</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
<b>B1</b>	5.098	7.312	8.430	3.958	0.000	5.970	16.703	14.557	14.825	13.320	15.354	14.378
<b>B2</b>	7.491	9.705	10.867	2.616	5.970	0.000	14.266	12.254	12.746	15.247	16.529	15.809
<b>M</b>	12.924	13.707	14.646	16.122	16.703	14.266	0.000	2.326	1.923	17.964	18.512	17.933
<b>M1</b>	11.270	12.634	13.349	13.797	14.557	12.254	2.326	0.000	0.626	17.155	17.967	17.419
<b>M2</b>	11.203	12.388	13.394	14.244	14.825	12.746	1.923	0.626	0.000	16.932	17.945	17.285
<b>E</b>	8.592	6.171	5.292	14.641	13.320	15.247	17.964	17.155	16.932	0.000	4.389	4.031
<b>E1</b>	11.537	9.050	8.480	16.256	15.354	16.529	18.512	17.967	17.945	4.389	0.000	1.138
<b>E2</b>	10.554	8.318	7.826	15.295	14.378	15.809	17.933	17.419	17.285	4.031	1.138	0.000

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

	<b>H</b>	<b>H1</b>	<b>H2</b>	<b>B</b>	<b>B1</b>	<b>B2</b>	<b>M</b>	<b>M1</b>	<b>M2</b>	<b>E</b>	<b>E1</b>	<b>E2</b>
<b>H</b>	0.000	0.650	0.656	10.207	10.393	9.704	12.611	12.216	11.729	41.743	33.650	33.528
<b>H1</b>	0.650	0.000	1.131	8.092	8.278	7.779	8.917	8.852	8.484	34.046	29.074	28.982
<b>H2</b>	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>B</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
<b>B1</b>	10.393	8.278	6.656	6.831	0.000	11.703	30.425	24.194	23.312	64.556	46.384	46.188
<b>B2</b>	9.704	7.779	6.159	6.683	11.703	0.000	22.666	18.113	17.199	63.969	45.943	45.748
<b>M</b>	12.611	8.917	9.428	29.218	30.425	22.666	0.000	0.617	0.612	60.900	44.199	44.013
<b>M1</b>	12.216	8.852	9.352	22.346	24.194	18.113	0.617	0.000	1.065	58.252	43.168	42.993
<b>M2</b>	11.729	8.484	8.986	21.408	23.312	17.199	0.612	1.065	0.000	58.239	43.139	42.963
<b>E</b>	41.743	34.046	35.161	65.138	64.556	63.969	60.900	58.252	58.239	0.000	0.250	0.251
<b>E1</b>	33.650	29.074	30.060	46.483	46.384	45.943	44.199	43.168	43.139	0.250	0.000	0.434
<b>E2</b>	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 XX. Values of  $c'$  for histograms drawn from the sizes of the known words.

	<b>H</b>	<b>H1</b>	<b>H2</b>	<b>B</b>	<b>B1</b>	<b>B2</b>	<b>M</b>	<b>M1</b>	<b>M2</b>	<b>E</b>	<b>E1</b>	<b>E2</b>
<b>H</b>	0.000	3.011	2.550	11.538	10.800	9.816	7.692	6.373	6.082	21.509	21.551	21.525
<b>H1</b>	3.011	0.000	0.882	11.888	11.567	11.030	9.136	8.219	8.398	19.255	19.297	19.270
<b>H2</b>	2.550	0.882	0.000	11.261	10.624	10.132	8.343	7.426	7.583	19.700	19.741	19.715
<b>B</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
<b>B1</b>	10.800	11.567	10.624	1.632	0.000	1.140	4.964	5.613	6.507	22.338	22.361	22.334
<b>B2</b>	9.816	11.030	10.132	1.878	1.140	0.000	4.092	4.629	5.434	22.338	22.361	22.334
<b>M</b>	7.692	9.136	8.343	5.724	4.964	4.092	0.000	1.722	1.901	22.315	22.338	22.312
<b>M1</b>	6.373	8.219	7.426	6.283	5.613	4.629	1.722	0.000	1.207	22.292	22.334	22.307
<b>M2</b>	6.082	8.398	7.583	7.133	6.507	5.434	1.901	1.207	0.000	22.315	22.338	22.312
<b>E</b>	21.509	19.255	19.700	22.338	22.338	22.338	22.315	22.292	22.315	0.000	2.472	3.751
<b>E1</b>	21.551	19.297	19.741	22.361	22.361	22.361	22.338	22.334	22.338	2.472	0.000	1.465
<b>E2</b>	21.525	19.270	19.715	22.338	22.334	22.334	22.312	22.307	22.312	3.751	1.465	0.000

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

	<b>H</b>	<b>H1</b>	<b>H2</b>	<b>B</b>	<b>B1</b>	<b>B2</b>	<b>M</b>	<b>M1</b>	<b>M2</b>	<b>E</b>	<b>E1</b>	<b>E2</b>
<b>H</b>	0.000	4.327	5.241	7.849	6.015	7.558	8.832	6.909	6.529	20.951	20.970	20.917
<b>H1</b>	4.327	0.000	1.026	11.322	9.668	11.188	11.859	10.383	10.227	16.623	16.642	16.589
<b>H2</b>	5.241	1.026	0.000	11.663	10.030	11.596	12.177	10.724	10.567	15.710	15.729	15.676
<b>B</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
<b>B1</b>	6.015	9.668	10.030	2.974	0.000	2.504	4.942	3.130	2.795	22.315	22.334	22.281
<b>B2</b>	7.558	11.188	11.596	1.476	2.504	0.000	2.639	1.789	1.655	22.338	22.334	22.281
<b>M</b>	8.832	11.859	12.177	2.996	4.942	2.639	0.000	2.281	2.661	22.338	22.334	22.281
<b>M1</b>	6.909	10.383	10.724	2.929	3.130	1.789	2.281	0.000	0.738	22.315	22.334	22.281
<b>M2</b>	6.529	10.227	10.567	2.862	2.795	1.655	2.661	0.738	0.000	22.315	22.334	22.281
<b>E</b>	20.951	16.623	15.710	22.315	22.315	22.338	22.338	22.315	22.315	0.000	4.870	6.237
<b>E1</b>	20.970	16.642	15.729	22.334	22.334	22.334	22.334	22.334	22.334	4.870	0.000	1.497
<b>E2</b>	20.917	16.589	15.676	22.281	22.281	22.281	22.281	22.281	22.281	6.237	1.497	0.000

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

	<b>H</b>	<b>H1</b>	<b>H2</b>	<b>B</b>	<b>B1</b>	<b>B2</b>	<b>M</b>	<b>M1</b>	<b>M2</b>	<b>E</b>	<b>E1</b>	<b>E2</b>
<b>H</b>	0.000	0.835	0.847	10.183	10.950	9.075	4.219	4.131	3.858	28.322	21.836	21.139
<b>H1</b>	0.835	0.000	1.456	8.314	8.916	7.563	4.081	4.064	3.857	24.599	19.810	19.248
<b>H2</b>	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>B</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
<b>B1</b>	10.950	8.916	6.810	3.777	0.000	6.571	15.289	11.548	12.190	39.275	28.450	27.424
<b>B2</b>	9.075	7.563	5.472	3.811	6.571	0.000	10.935	8.189	8.804	38.128	27.622	26.624
<b>M</b>	4.219	4.081	2.055	14.417	15.289	10.935	0.000	0.422	0.416	34.862	25.388	24.474
<b>M1</b>	4.131	4.064	2.100	10.427	11.548	8.189	0.422	0.000	0.725	34.184	25.154	24.267
<b>M2</b>	3.858	3.857	1.893	11.101	12.190	8.804	0.416	0.725	0.000	34.032	25.032	24.149
<b>E</b>	28.322	24.599	25.815	38.938	39.275	38.128	34.862	34.184	34.032	0.000	0.382	0.410
<b>E1</b>	21.836	19.810	20.821	28.127	28.450	27.622	25.388	25.154	25.032	0.382	0.000	0.686
<b>E2</b>	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 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 leaves. 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
<b>S1</b>	recorded 'front center'	68545
<b>W1-1</b>	first wavelet approximation	31
<b>W2-1</b>	higher wavelet leaf	17147
<b>S2</b>	recorded 'front left'	71042
<b>W1-2</b>	first wavelet approximation	32
<b>W2-2</b>	higher wavelet leaf	17771
<b>S3</b>	recorded 'rear center'	65026
<b>W1-3</b>	first wavelet approximation	30
<b>W2-3</b>	higher wavelet leaf	16267
<b>S4</b>	recorded 'rear left'	63010
<b>W1-4</b>	first wavelet approximation	30
<b>W2-4</b>	higher wavelet leaf	15763
<b>S5</b>	noise	67579
<b>W1-5</b>	first wavelet approximation	31
<b>W2-5</b>	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.

	<b>S1</b>	<b>W1-1</b>	<b>W2-1</b>	<b>S2</b>	<b>W1-2</b>	<b>W2-2</b>	<b>S3</b>	<b>W1-3</b>	<b>W2-3</b>	<b>S4</b>	<b>W1-4</b>	<b>W2-4</b>	<b>S5</b>	<b>W1-5</b>	<b>W2-5</b>
<b>S1</b>	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
<b>W1-1</b>	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
<b>W2-1</b>	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
<b>S2</b>	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
<b>W1-2</b>	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
<b>W2-2</b>	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
<b>S3</b>	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
<b>W1-3</b>	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
<b>W2-3</b>	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
<b>S4</b>	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
<b>W1-4</b>	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
<b>W2-4</b>	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
<b>S5</b>	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
<b>W1-5</b>	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
<b>W2-5</b>	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

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
<b>Pale</b>	Sanctus 69 from G. P. da Palestrina	719
<b>Bach1</b>	BWV735 from J. S. Bach	236
<b>Bach2</b>	BWV648 from J. S. Bach	272
<b>Moza1</b>	K80 from W. A. Mozart	538
<b>Moza2</b>	K458 from W. A. Mozart	4218
<b>Beet1</b>	Opus 18, n1, mov. 3 from L. van Beethoven	1289
<b>Beet2</b>	Opus 132 from L. van Beethoven	17884
<b>Schön</b>	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 were chosen to reflect music history timeline. Works by the same composer were chosen among the first and last 10% of all he produced.

	<b>Pale</b>	<b>Bach1</b>	<b>Bach2</b>	<b>Moza1</b>	<b>Moza2</b>	<b>Beet1</b>	<b>Beet2</b>	<b>Schön</b>
<b>Pale</b>	0.00	1.36	1.39	2.05	2.44	1.72	2.33	1.89
<b>Bach1</b>	1.36	0.00	0.88	1.00	1.60	1.02	1.01	1.54
<b>Bach2</b>	1.39	0.88	0.00	1.73	1.22	1.41	1.68	1.52
<b>Moza1</b>	2.05	1.00	1.73	0.00	2.14	1.65	1.79	1.79
<b>Moza2</b>	2.44	1.60	1.22	2.14	0.00	1.62	4.20	1.84
<b>Beet1</b>	1.72	1.02	1.41	1.65	1.62	0.00	2.34	2.03
<b>Beet2</b>	2.33	1.01	1.68	1.79	4.20	2.34	0.00	2.02
<b>Schön</b>	1.89	1.54	1.52	1.79	1.84	2.03	2.02	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
<b>cpu1</b>	workload of the most active processor	888
<b>cpu2</b>	workload of the second most active processor	422
<b>cpu3</b>	workload of the third most active processor	1046
<b>mem</b>	RAM use in kB	557
<b>p1</b>	workload use of most consuming process	1197
<b>m1</b>	RAM use of most consuming process	1197
<b>p2</b>	workload use of second most consuming process	1197
<b>m2</b>	RAM use of second most consuming process	1197
<b>p3</b>	RAM use of third most consuming process	1197
<b>m3</b>	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.

	<b>cpu1</b>	<b>cpu2</b>	<b>cpu3</b>	<b>mem</b>	<b>p1</b>	<b>m1</b>	<b>p2</b>	<b>m2</b>	<b>p3</b>	<b>m3</b>
<b>cpu1</b>	0.00	5.13	4.73	4.84	8.57	11.03	1.84	9.25	3.39	9.94
<b>cpu2</b>	5.13	0.00	2.97	3.83	8.89	9.90	4.72	9.03	3.64	9.00
<b>cpu3</b>	4.73	2.97	0.00	3.69	8.63	13.44	4.63	12.50	4.51	12.00
<b>mem</b>	4.84	3.83	3.69	0.00	4.16	10.77	4.60	9.53	3.97	9.48
<b>p1</b>	8.57	8.89	8.63	4.16	0.00	13.75	10.08	12.77	9.85	12.53
<b>m1</b>	11.03	9.90	13.44	10.77	13.75	0.00	12.26	12.53	10.57	15.43
<b>p2</b>	1.84	4.72	4.63	4.60	10.08	12.26	0.00	10.51	2.82	10.85
<b>m2</b>	9.25	9.03	12.50	9.53	12.77	12.53	10.51	0.00	10.30	8.67
<b>p3</b>	3.39	3.64	4.51	3.97	9.85	10.57	2.82	10.30	0.00	10.30
<b>m3</b>	9.94	9.00	12.00	9.48	12.53	15.43	10.85	8.67	10.30	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<sup>3</sup>. 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.

<sup>1</sup>R. Chicheportiche and J.-P. Bouchaud, “Weighted kolmogorov-smirnov test: Accounting for the tails,” *Physical Review E* **86**, 041115 (2012).

<sup>2</sup>G. Deleuze, *Difference and repetition* (Columbia University Press, 1994).

<sup>3</sup>R. Fabbri, “Gmane python package for analysing public email lists,” <https://pypi.python.org/pypi/gmane> (2015).