Distances between histograms

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: 28 October 2015)

This document presents reference values for a distance metric derived from the Kolmogorov-Smirnov statistical test. Each measure is a distance between two histograms. The sections are self-explanatory on deriving benchmarks by comparing samples from usual distributions and on exemplifying the power of the acquired knowledge.

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

I. INTRODUCTION

Be $F_{1,n}$ and $F_{2,n'}$ 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:

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

where $D_{n,n'} = sup_x[F_{1,n} - F_{2,n'}]$ and $c(\alpha)$ are related to the critical region 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, $D_{n,n'}$ is given as are n and n'. All terms in equation 1 are positive and $c(\alpha)$ can be isolated:

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

When c' is high, low values of α favor rejecting the null hypothesis. For example, when c' is greater than ≈ 1.7, one might assume that $F_{1,n}$ and $F_{2,n'}$ are outcomes of different distributions. More importantly for us is that c' is a measure of distance between both distributions[1]. The main contribution of the following sections is the explicit display of reference values from which one might derive knowledge from collections of empirical measures of c' or even of a single value of c'.

A. Philosophical and technological note

Difference and equivalence is of central role in human cognition, philosophy and science. This fact is so deeply recognized that thinkers often reduce thought to classifications, e.g. through the mathematical concept of equivalence classes[? ?]. Histograms are very immediate and informative 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 minimum delivered by the following tables is that the software implementation that render them present measures of how different are the underlying distributions of given samples.

B. Document outline

Section II expose reference values drawn from simulations. Section III exemplifies the use of such reference values to make sense of phenomena. Section VII hold final remarks. Software and data specification are given in Appendix A.

II. REFERENCES THROUGH SIMULATIONS

On every case, values of c' are given for simulations involving at least normal, uniform, triangular, Weibull and power-law distributions.

A. When the null hypothesis is true

If the null hypothesis is true, than the number of rejections of the null hypothesis (that is: $c' > c(\alpha)$) 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.

The most important result of this section is that $|C'(\alpha)|$ rarely exceeds αN_c, where N_c is the number of

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

comparisons made, no matter what probability distribution type or the specific setting. Also important are that $c' > c(\alpha)$ in many cases and that α is a good estimate of the upper limit of the probability 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	6	2	4
5.0	0.050	1.36	3	0	3
2.5	0.025	1.48	1	0	1
1.0	0.010	1.63	1	0	0
0.5	0.005	1.73	0	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 critical values of α. The number of comparisons is $N_c = 100$, each with the sample size of $N_o = 1000$ observations. Each histogram have $N_b = 30$ equally spaced bins. The N_o values of c_1' were calculated using simulations of normal distributions with $\mu = 0$ and $\sigma = 1$. The N_o values of c_2' were calculated using simulations of normal distributions with $\mu = 3$ and $\sigma = 2$. The N_o values of c_3' were calculated using simulations of normal distributions with $\mu = 6$ and $\sigma = 3$. Over all N_c comparisons, $\mu(c_1') = 0.7240$ and $\sigma(c_1') = 0.2850$, $\mu(c_2') = 0.7079$ and $\sigma(c_2') = 0.2314$, $\mu(c_3') = 0.7187$ and $\sigma(c_3') = 0.2559$.

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	10	6
5.0	0.050	1.36	6	3	2
2.5	0.025	1.48	2	2	1
1.0	0.010	1.63	1	0	1
0.5	0.005	1.73	0	0	1
0.1	0.001	1.95	0	0	0

TABLE II. The theoretical maximum number αN_c of rejections of the null hypothesis for critical values of α. The number of comparisons is $N_c = 100$, each with the sample size of $N_o = 1000$ observations. Each histogram have $N_b = 30$ equally spaced bins. The N_o values of c_1' were calculated using simulations of uniform distributions within $[0, 1)$. The N_o values of c_2' were calculated using simulations of uniform distributions within $[2, 6)$. The N_o values of c_3' were calculated using simulations of normal distributions with $\mu = 4$ and $\sigma = 10$. Over all N_c comparisons, $\mu(c_1') = 0.7965$ and $\sigma(c_1') = 0.2768$, $\mu(c_2') = 0.7773$ and $\sigma(c_2') = 0.2657$, $\mu(c_3') = 0.7828$ and $\sigma(c_3') = 0.2589$.

B. When the null hypothesis if false

The $m(c')$ are median values of c'. $\overline{C'(\alpha)} = \frac{|C'(\alpha)|}{N_c}$ is the fraction of rejection of the null hypothesis with critical region α.

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	8	3	10
5.0	0.050	1.36	0	5	3	7
2.5	0.025	1.48	0	3	3	4
1.0	0.010	1.63	0	1	1	3
0.5	0.005	1.73	0	0	0	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 number of comparisons is $N_c = 100$, each with the sample size of $N_o = 1000$ observations. Each histogram have $N_b = 30$ equally spaced bins. The N_o values of c_1' were calculated using simulations of 1-parameter Weibull distributions with $a = 0.1$. The N_o values of c_2' were calculated using simulations of 1-parameter Weibull distributions with $a = 2$. The N_o values of c_3' 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, $\mu(c_1') = 0.0646$ and $\sigma(c_1') = 0.0373$, $\mu(c_2') = 0.7737$ and $\sigma(c_2') = 0.2842$, $\mu(c_3') = 0.6977$ and $\sigma(c_3') = 0.2629$. $\mu(c_4') = 0.7782$ and $\sigma(c_4') = 0.3116$.

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	4	10	7	9	1
5.0	0.050	1.36	3	3	4	5	0
2.5	0.025	1.48	1	1	2	3	0
1.0	0.010	1.63	1	0	1	2	0
0.5	0.005	1.73	0	0	1	1	0
0.1	0.001	1.95	0	0	0	0	0

TABLE IV. The theoretical maximum number αN_c of rejections of the null hypothesis for critical values of α. The number of comparisons is $N_c = 100$, each with the sample size of $N_o = 1000$ observations. Each histogram have $N_b = 30$ equally spaced bins. The N_o values of c_1' were calculated using simulations of power functions distributions with $a = 0.3$. The N_o values of c_2' were calculated using simulations of power functions distributions with $a = 1$. The N_o values of c_3' were calculated using simulations of power functions distributions with $a = 2$. The N_o values of c_4' were calculated using simulations of power functions distributions with $a = 3$. The N_o values of c_5' were calculated using simulations of power functions distributions with $a = 4$. Over all N_c comparisons, $\mu(c_1') = 0.7343$ and $\sigma(c_1') = 0.2654$, $\mu(c_2') = 0.8139$ and $\sigma(c_2') = 0.2698$, $\mu(c_3') = 0.7328$ and $\sigma(c_3') = 0.2731$. $\mu(c_4') = 0.7623$ and $\sigma(c_4') = 0.3053$. $\mu(c_5') = 0.7269$ and $\sigma(c_5') = 0.1949$.

III. EXAMPLE USES IN EMPIRICAL DATA

IV. TEXT

V. AUDIO

VI. OS STATUS

VII. CONCLUSIONS

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. The authors are grateful to the American Jewish Committee for maintaining an online copy of the Adorno book used on the epigraph[?] , to GMANE creators and maintainers for the public email list data, to the communities of the email lists and other groups used in the analysis, and to the Brazilian Presidency of the Republic for keeping Participabr code and data open. We are also grateful to developers and users of Python scientific tools.

Appendix A: Software and data specifications

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

σ	$\mu(c')$	$\sigma(c')$	m(c')	min(c')	max(c')	$\overline{C'(0.1)}$	$\overline{C'(0.05)}$	$\overline{C'(0.025)}$	$\overline{C'(0.01)}$	$\overline{C'(0.005)}$	$\overline{C'(0.001)}$
0.5	3.843	0.281	3.790	3.153,3.265,3.332	4.450,4.539,4.696	1.000	1.000	1.000	1.000	1.000	1.000
0.6	2.965	0.269	2.929	2.326,2.437,2.437	3.578,3.578,3.578	1.000	1.000	1.000	1.000	1.000	1.000
0.7	2.146	0.263	2.135	1.565,1.632,1.655	2.728,2.773,2.907	1.000	1.000	1.000	0.990	0.970	0.730
0.8	1.587	0.286	1.543	0.984,1.051,1.073	2.214,2.258,2.370	0.920	0.790	0.630	0.370	0.250	0.140
0.9	0.976	0.256	0.973	0.447,0.537,0.559	1.543,1.588,1.632	0.180	0.090	0.040	0.010	0.000	0.000
1.0	0.729	0.295	0.671	0.246,0.291,0.313	1.364,1.677,1.766	0.070	0.030	0.020	0.020	0.010	0.000
1.1	0.959	0.263	0.906	0.514,0.537,0.559	1.588,1.588,1.789	0.150	0.090	0.050	0.010	0.010	0.000
1.2	1.308	0.241	1.319	0.738,0.872,0.872	1.789,1.878,2.102	0.630	0.430	0.200	0.090	0.050	0.010
1.3	1.735	0.266	1.688	0.939,1.118,1.140	2.214,2.281,2.303	0.970	0.950	0.820	0.640	0.460	0.240
1.4	2.068	0.299	2.035	1.498,1.588,1.588	2.773,2.773,3.533	1.000	1.000	1.000	0.970	0.910	0.610
1.5	2.440	0.248	2.381	1.945,1.968,2.035	2.929,2.929,2.974	1.000	1.000	1.000	1.000	1.000	0.990
1.6	2.814	0.259	2.817	2.236,2.326,2.348	3.354,3.376,3.578	1.000	1.000	1.000	1.000	1.000	1.000
1.7	3.042	0.313	3.041	2.281,2.504,2.504	3.757,3.824,3.846	1.000	1.000	1.000	1.000	1.000	1.000
1.8	3.410	0.290	3.388	2.706,2.750,2.840	3.935,4.003,4.338	1.000	1.000	1.000	1.000	1.000	1.000
1.9	3.583	0.250	3.578	2.840,2.974,2.996	4.114,4.137,4.181	1.000	1.000	1.000	1.000	1.000	1.000
2.0	3.806	0.289	3.779	3.265,3.287,3.309	4.405,4.562,4.651	1.000	1.000	1.000	1.000	1.000	1.000

TABLE V. Location and dispersion of $N_c = 100$ measurements of c' through simulations with normal distributions and $N_o = 1000$ events each. $N_b = 30$ equal bins were used to make the histograms. One normal distribution is fixed, with $\mu = 0$ and $\sigma = 1$, and compared agaist normal distributions with $\mu = 0$ and different values of σ.

μ	$\mu(c')$	$\sigma(c')$	m(c')	min(c')	max(c')	$\overline{C'(0.1)}$	$\overline{C'(0.05)}$	$\overline{C'(0.025)}$	$\overline{C'(0.01)}$	$\overline{C'(0.005)}$	$\overline{C'(0.001)}$
0.0	0.709	0.260	0.693	0.268,0.313,0.313	1.364,1.386,1.386	0.050	0.030	0.000	0.000	0.000	0.000
0.1	1.258	0.406	1.275	0.380,0.425,0.470	1.990,2.124,2.258	0.540	0.420	0.280	0.200	0.130	0.040
0.2	1.992	0.441	2.001	0.850,1.185,1.207	3.108,3.108,3.265	0.970	0.920	0.890	0.810	0.730	0.540
0.3	2.836	0.423	2.829	1.834,1.968,1.968	3.690,3.734,3.824	1.000	1.000	1.000	1.000	1.000	0.990
0.4	3.710	0.482	3.667	2.706,2.840,2.996	4.763,5.143,5.322	1.000	1.000	1.000	1.000	1.000	1.000
0.5	4.527	0.476	4.539	3.309,3.309,3.555	5.456,5.545,5.635	1.000	1.000	1.000	1.000	1.000	1.000
0.6	5.325	0.456	5.344	4.137,4.338,4.494	6.350,6.440,6.485	1.000	1.000	1.000	1.000	1.000	1.000
0.7	6.145	0.439	6.127	5.210,5.232,5.367	7.044,7.066,7.111	1.000	1.000	1.000	1.000	1.000	1.000
0.8	7.028	0.431	7.010	5.680,6.194,6.283	7.938,8.005,8.117	1.000	1.000	1.000	1.000	1.000	1.000
0.9	7.841	0.451	7.815	6.909,6.909,6.954	8.810,8.810,8.900	1.000	1.000	1.000	1.000	1.000	1.000
1.0	8.562	0.499	8.542	7.468,7.513,7.513	9.503,9.705,9.772	1.000	1.000	1.000	1.000	1.000	1.000

TABLE VI. Location and dispersion of $N_c = 100$ measurements of c' through simulations with normal distributions and $N_o = 1000$ events each. $N_b = 30$ equal bins were used to make the histograms. One normal distribution is fixed, with $\mu = 0$ and $\sigma = 1$, and compared agaist normal distributions with different values of μ and fixed $\sigma = 1$.

b	$\mu(c')$	$\sigma(c')$	m(c')	min(c')	max(c')	$\overline{C'(0.1)}$	$\overline{C'(0.05)}$	$\overline{C'(0.025)}$	$\overline{C'(0.01)}$	$\overline{C'(0.005)}$	$\overline{C'(0.001)}$
0.7	6.708	0.347	6.697	5.836,5.926,5.993	7.379,7.401,7.513	1.000	1.000	1.000	1.000	1.000	1.000
0.75	5.448	0.268	5.434	4.897,4.964,4.986	5.970,6.015,6.127	1.000	1.000	1.000	1.000	1.000	1.000
0.8	4.487	0.317	4.461	3.846,3.913,3.935	5.210,5.277,5.322	1.000	1.000	1.000	1.000	1.000	1.000
0.85	3.305	0.255	3.287	2.683,2.750,2.795	3.824,3.868,3.891	1.000	1.000	1.000	1.000	1.000	1.000
0.9	2.330	0.246	2.359	1.610,1.878,1.878	2.795,2.885,2.907	1.000	1.000	1.000	0.990	0.990	0.930
0.95	1.306	0.295	1.263	0.716,0.805,0.850	2.035,2.057,2.080	0.540	0.350	0.290	0.130	0.090	0.040
1.0	0.756	0.259	0.716	0.268,0.313,0.380	1.297,1.386,1.632	0.080	0.020	0.010	0.010	0.000	0.000
1.05	1.230	0.259	1.241	0.716,0.760,0.805	1.744,1.856,1.856	0.500	0.360	0.160	0.060	0.030	0.000
1.1	2.101	0.262	2.068	1.431,1.610,1.610	2.706,2.862,2.862	1.000	1.000	0.990	0.970	0.950	0.680
1.15	2.965	0.257	2.940	2.482,2.549,2.594	3.600,3.712,3.734	1.000	1.000	1.000	1.000	1.000	1.000
1.2	3.752	0.270	3.779	3.153,3.220,3.220	4.271,4.316,4.472	1.000	1.000	1.000	1.000	1.000	1.000
1.25	4.467	0.280	4.427	3.891,3.980,4.003	5.009,5.031,5.434	1.000	1.000	1.000	1.000	1.000	1.000
1.3	5.161	0.318	5.176	4.383,4.405,4.494	5.724,5.836,5.970	1.000	1.000	1.000	1.000	1.000	1.000
1.35	5.767	0.325	5.758	4.986,5.054,5.098	6.395,6.641,6.775	1.000	1.000	1.000	1.000	1.000	1.000

TABLE VII. Location and dispersion of $N_c = 100$ measurements of c' through simulations with uniform distributions and $N_o = 1000$ events each. $N_b = 30$ equal bins were used to make the histograms. One uniform distribution has the fixed domain $[0, 1)$. The other uniform distribution in each comparison is also centered around 0.5, but spread over $b = b_u - b_l$ there b_l and b_u are the lower and upper boudaries.

μ	$\mu(c')$	$\sigma(c')$	m(c')	min(c')	max(c')	$\overline{C'(0.1)}$	$\overline{C'(0.05)}$	$\overline{C'(0.025)}$	$\overline{C'(0.01)}$	$\overline{C'(0.005)}$	$\overline{C'(0.001)}$
0.5	0.778	0.283	0.727	0.224,0.402,0.425	1.521,1.811,1.834	0.060	0.050	0.030	0.020	0.020	0.000
0.55	1.632	0.314	1.610	1.051,1.051,1.073	2.326,2.393,2.437	0.930	0.810	0.660	0.490	0.320	0.160
0.6	2.728	0.292	2.683	2.191,2.214,2.236	3.376,3.399,3.511	1.000	1.000	1.000	1.000	1.000	1.000
0.65	3.840	0.334	3.824	3.086,3.175,3.220	4.517,4.562,4.584	1.000	1.000	1.000	1.000	1.000	1.000
0.7	5.040	0.330	5.009	4.383,4.450,4.494	5.814,5.836,5.836	1.000	1.000	1.000	1.000	1.000	1.000
0.75	6.091	0.291	6.082	5.255,5.613,5.613	6.641,6.731,6.798	1.000	1.000	1.000	1.000	1.000	1.000
0.8	7.114	0.381	7.111	6.261,6.440,6.485	8.072,8.072,8.095	1.000	1.000	1.000	1.000	1.000	1.000
0.85	8.291	0.358	8.296	7.424,7.446,7.670	8.967,8.967,9.034	1.000	1.000	1.000	1.000	1.000	1.000
0.9	9.398	0.355	9.391	8.609,8.743,8.743	10.152,10.241,10.375	1.000	1.000	1.000	1.000	1.000	1.000
0.95	10.416	0.326	10.375	9.705,9.772,9.772	11.136,11.136,11.315	1.000	1.000	1.000	1.000	1.000	1.000
1.0	11.603	0.312	11.605	10.688,10.934,11.024	12.298,12.298,12.455	1.000	1.000	1.000	1.000	1.000	1.000
1.05	12.655	0.300	12.634	11.874,12.097,12.164	13.238,13.305,13.416	1.000	1.000	1.000	1.000	1.000	1.000
1.1	13.729	0.343	13.729	12.768,12.992,13.103	14.624,14.691,14.915	1.000	1.000	1.000	1.000	1.000	1.000

TABLE VIII. Location and dispersion of $N_c = 100$ measurements of c' through simulations with uniform distributions and $N_o = 1000$ events each. $N_b = 30$ equal bins were used to make the histograms. One uniform distribution has the fixed domain $[0, 1)$. The other uniform distribution in each comparison have varied mean values but always spread over $b = b_u - b_l$ there b_l and b_u are the lower and upper boudaries.

a	$\mu(c')$	$\sigma(c')$	m(c')	min(c')	max(c')	$\overline{C'(0.1)}$	$\overline{C'(0.05)}$	$\overline{C'(0.025)}$	$\overline{C'(0.01)}$	$\overline{C'(0.005)}$	$\overline{C'(0.001)}$
0.01	0.027	0.011	0.022	0.022,0.022,0.022	0.067,0.067,0.067	0.000	0.000	0.000	0.000	0.000	0.000
0.1	0.162	0.094	0.157	0.022,0.022,0.022	0.380,0.425,0.514	0.000	0.000	0.000	0.000	0.000	0.000
0.3	1.758	0.698	1.778	0.358,0.358,0.402	3.287,3.332,3.332	0.790	0.730	0.640	0.560	0.530	0.340
0.5	3.967	0.409	4.003	2.728,3.063,3.086	4.562,4.629,5.009	1.000	1.000	1.000	1.000	1.000	1.000
0.7	3.502	0.503	3.477	2.571,2.639,2.795	4.584,4.785,5.031	1.000	1.000	1.000	1.000	1.000	1.000
0.9	2.987	0.400	2.963	2.102,2.147,2.258	3.757,3.757,3.935	1.000	1.000	1.000	1.000	1.000	1.000
1.1	1.982	0.372	1.912	1.118,1.252,1.364	2.929,2.996,3.041	0.990	0.980	0.940	0.870	0.720	0.470
1.3	1.182	0.346	1.163	0.514,0.559,0.581	1.856,2.303,2.504	0.410	0.300	0.150	0.090	0.050	0.020
1.5	0.688	0.247	0.682	0.291,0.291,0.291	1.319,1.476,1.476	0.040	0.020	0.000	0.000	0.000	0.000
1.7	1.057	0.252	1.073	0.447,0.559,0.626	1.476,1.610,2.057	0.260	0.120	0.020	0.010	0.010	0.010
1.9	1.619	0.338	1.576	0.894,1.029,1.051	2.527,2.549,2.594	0.900	0.830	0.620	0.440	0.350	0.130
2.1	2.187	0.346	2.169	1.498,1.543,1.588	3.108,3.108,3.265	1.000	1.000	1.000	0.970	0.940	0.730
2.3	2.623	0.374	2.560	1.968,1.990,2.057	3.466,3.466,3.667	1.000	1.000	1.000	1.000	1.000	1.000
2.5	3.033	0.409	2.963	2.258,2.326,2.348	4.003,4.003,4.114	1.000	1.000	1.000	1.000	1.000	1.000
2.7	3.486	0.369	3.488	2.504,2.795,2.817	4.159,4.249,4.316	1.000	1.000	1.000	1.000	1.000	1.000
2.9	3.968	0.398	3.947	2.929,3.153,3.198	4.718,4.763,5.188	1.000	1.000	1.000	1.000	1.000	1.000

TABLE IX. Location and dispersion of $N_c = 100$ measurements of c' through simulations with 1-parameter Weibull distributions and $N_o = 1000$ events each. $N_b = 30$ equal bins were used to make the histograms. One Weibull distribution has the fixed shape parameter $a = 1.5$. The other Weibull distribution in each comparison has varied values of a.

a	$\mu(c')$	$\sigma(c')$	m(c')	min(c')	max(c')	$\overline{C'(0.1)}$	$\overline{C'(0.05)}$	$\overline{C'(0.025)}$	$\overline{C'(0.01)}$	$\overline{C'(0.005)}$	$\overline{C'(0.001)}$
0.7	6.273	0.449	6.295	4.942,5.143,5.165	7.044,7.088,7.491	1.000	1.000	1.000	1.000	1.000	1.000
0.9	4.376	0.486	4.349	3.265,3.287,3.444	5.344,5.367,5.478	1.000	1.000	1.000	1.000	1.000	1.000
1.1	2.802	0.477	2.817	1.655,1.811,1.901	3.757,3.913,4.181	1.000	1.000	1.000	1.000	0.990	0.960
1.3	1.464	0.439	1.465	0.447,0.470,0.559	2.393,2.437,2.571	0.690	0.570	0.470	0.370	0.270	0.130
1.5	0.763	0.284	0.671	0.291,0.380,0.380	1.453,1.677,1.699	0.100	0.050	0.020	0.020	0.000	0.000
1.7	1.382	0.349	1.386	0.537,0.559,0.648	2.124,2.191,2.326	0.650	0.550	0.360	0.230	0.140	0.070
1.9	2.233	0.425	2.225	1.342,1.386,1.476	3.242,3.265,3.376	1.000	0.990	0.970	0.960	0.900	0.690
2.1	3.022	0.456	3.019	2.169,2.191,2.214	3.935,4.047,4.427	1.000	1.000	1.000	1.000	1.000	1.000
2.3	3.677	0.441	3.779	2.370,2.393,2.706	4.472,4.517,4.785	1.000	1.000	1.000	1.000	1.000	1.000
2.5	4.340	0.451	4.394	3.332,3.332,3.399	5.076,5.165,5.210	1.000	1.000	1.000	1.000	1.000	1.000

TABLE X. Location and dispersion of $N_c = 100$ measurements of c' through simulations with power function distributions and $N_o = 1000$ events each. $N_b = 30$ equal bins were used to make the histograms. 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.

label	description	chars	tokens	sents	$\mu(kw)$	$\sigma(kw)$	$\mu(sw)$	$\sigma(sw)$
Hamlet	Hamlet by Shakespeare	162881	37360	3106	3.549	1.762	2.721	1.011
Bible	King James Version of the Holly Bible	4332554	1010654	30103	3.745	1.711	2.927	1.044
Moby	Moby Dick by Herman Melville	1242990	260819	10059	4.105	2.184	2.847	1.096
Memórias	Esaú e Jacó from Machado de Assis	355706	77098	3822	2.186	1.376	1.486	0.502

TABLE XI. General description of the texts used to exemplify the use of the c' values. 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 used are used in next table for comparison through c'. This table holds has the only purpose of contextualizing next table.

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

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

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

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

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

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

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

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

	$S1$	$W_1 1$	$W_2 1$	$S2$	$W_1 2$	$W_2 2$	$S3$	$W_1 3$	$W_2 3$	$S4$	$W_1 4$	$W_2 4$	$S5$	$W_1 5$	$W_2 5$
$S1$	0.000	0.638	1.788	0.736	2.121	1.788	0.579	2.657	1.788	1.956	2.599	1.788	1.778	1.129	1.788
$W_1 1$	0.638	0.000	22.327	1.138	2.867	22.327	0.851	2.242	22.327	2.858	4.600	22.327	2.251	29.901	22.327
$W_2 1$	1.788	22.327	0.000	2.062	20.747	0.000	2.054	19.074	0.000	1.968	22.221	0.000	1.702	19.829	0.000
$S2$	0.736	1.138	2.062	0.000	3.030	2.062	0.721	2.944	2.062	1.599	1.552	2.062	1.540	0.842	2.062
$W_1 2$	2.121	2.867	20.747	3.030	0.000	20.747	2.034	2.981	20.747	3.365	1.602	20.747	2.511	33.095	20.747
$W_2 2$	1.788	22.327	0.000	2.062	20.747	0.000	2.054	19.074	0.000	1.968	22.221	0.000	1.702	19.829	0.000
$S3$	0.579	0.851	2.054	0.721	2.034	2.054	0.000	3.090	2.054	1.936	2.244	2.054	1.763	1.208	2.054
$W_1 3$	2.657	2.242	19.074	2.944	2.981	19.074	3.090	0.000	19.074	2.901	3.176	19.074	2.258	30.109	19.074
$W_2 3$	1.788	22.327	0.000	2.062	20.747	0.000	2.054	19.074	0.000	1.968	22.221	0.000	1.702	19.829	0.000
$S4$	1.956	2.858	1.968	1.599	3.365	1.968	1.936	2.901	1.968	0.000	3.411	1.968	1.377	2.648	1.968
$W_1 4$	2.599	4.600	22.221	1.552	1.602	22.221	2.244	3.176	22.221	3.411	0.000	22.221	2.512	33.243	22.221
$W_2 4$	1.788	22.327	0.000	2.062	20.747	0.000	2.054	19.074	0.000	1.968	22.221	0.000	1.702	19.829	0.000
$S5$	1.778	2.251	1.702	1.540	2.511	1.702	1.763	2.258	1.702	1.377	2.512	1.702	0.000	2.407	1.702
$W_1 5$	1.129	29.901	19.829	0.842	33.095	19.829	1.208	30.109	19.829	2.648	33.243	19.829	2.407	0.000	19.829
$W_2 5$	1.788	22.327	0.000	2.062	20.747	0.000	2.054	19.074	0.000	1.968	22.221	0.000	1.702	19.829	0.000

TABLE XVI. Values of c' for histograms drawn from sound PCM samples and wavelet leaf coefficient.