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: 27 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:

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

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:

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

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.

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 IV 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 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.

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

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	2	5	1
5.0	0.050	1.36	0	3	0
2.5	0.025	1.48	0	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 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, $\mu(c'_1) = 0.6719$ and $\sigma(c'_1) = 0.2227$, $\mu(c'_2) = 0.7544$ and $\sigma(c'_2) = 0.2449$, $\mu(c'_3) = 0.6418$ and $\sigma(c'_3) = 0.2115$.

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	8	6
5.0	0.050	1.36	3	4	2
2.5	0.025	1.48	1	3	2
1.0	0.010	1.63	1	1	1
0.5	0.005	1.73	1	1	0
0.1	0.001	1.95	1	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, $\mu(c'_1) = 0.8010$ and $\sigma(c'_1) = 0.2810$, $\mu(c'_2) = 0.7471$ and $\sigma(c'_2) = 0.2921$, $\mu(c'_3) = 0.7612$ and $\sigma(c'_3) = 0.2608$.

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.

III. EXAMPLE USES IN EMPIRICAL DATA

IV. 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.

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	5	3	1
5.0	0.050	1.36	0	1	1	0
2.5	0.025	1.48	0	1	0	0
1.0	0.010	1.63	0	1	0	0
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, $\mu(c'_1) = 0.0684$ and $\sigma(c'_1) = 0.0494$, $\mu(c'_2) = 0.7395$ and $\sigma(c'_2) = 0.2480$, $\mu(c'_3) = 0.7256$ and $\sigma(c'_3) = 0.2406$. $\mu(c'_4) = 0.6831$ and $\sigma(c'_4) = 0.2130$.

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	3.859	0.269	3.812	3.287,3.354,3.421	4.383,4.405,4.562	1.000	1.000	1.000	1.000	1.000	1.000
0.6	3.004	0.308	3.019	2.281,2.326,2.393	3.578,3.891,3.935	1.000	1.000	1.000	1.000	1.000	1.000
0.7	2.226	0.308	2.225	1.453,1.677,1.699	2.862,2.907,3.242	1.000	1.000	0.990	0.990	0.950	0.820
0.8	1.542	0.279	1.509	0.827,1.029,1.029	2.191,2.191,2.236	0.880	0.760	0.520	0.330	0.250	0.080
0.9	1.011	0.266	1.006	0.559,0.581,0.604	1.610,1.699,1.878	0.200	0.130	0.050	0.020	0.010	0.000
1.0	0.714	0.217	0.693	0.335,0.335,0.358	1.163,1.207,1.453	0.010	0.010	0.000	0.000	0.000	0.000
1.1	0.911	0.270	0.894	0.425,0.470,0.492	1.409,1.431,1.699	0.130	0.070	0.010	0.010	0.000	0.000
1.2	1.331	0.255	1.319	0.760,0.783,0.872	1.834,1.856,1.856	0.660	0.420	0.290	0.160	0.060	0.000
1.3	1.737	0.289	1.766	0.984,1.006,1.073	2.303,2.370,2.482	0.950	0.920	0.800	0.670	0.510	0.210
1.4	2.113	0.310	2.102	1.521,1.543,1.565	2.795,2.795,2.885	1.000	1.000	1.000	0.940	0.900	0.660
1.5	2.454	0.255	2.415	1.766,1.834,1.878	3.041,3.063,3.086	1.000	1.000	1.000	1.000	1.000	0.960
1.6	2.771	0.298	2.795	2.169,2.191,2.191	3.421,3.488,3.645	1.000	1.000	1.000	1.000	1.000	1.000
1.7	3.086	0.303	3.063	2.370,2.460,2.460	3.645,3.667,3.891	1.000	1.000	1.000	1.000	1.000	1.000
1.8	3.345	0.307	3.309	2.728,2.728,2.795	4.047,4.159,4.360	1.000	1.000	1.000	1.000	1.000	1.000
1.9	3.561	0.236	3.555	3.086,3.108,3.153	4.070,4.070,4.181	1.000	1.000	1.000	1.000	1.000	1.000
2.0	3.808	0.291	3.779	3.220,3.242,3.287	4.472,4.517,4.942	1.000	1.000	1.000	1.000	1.000	1.000

TABLE IV. 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.