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:

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

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:

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

When $c'(\alpha)$ is high, low values of α favor rejecting the null hypothesis. For example, when $c'(\alpha)$ 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 $c'(\alpha)$ 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 $c'(\alpha)$ or even a single value of $c'(\alpha)$.

A. Philosophical and technological note

Difference and equivalence is of central role in human cognition, philosophy and science[? ?] . This is so deeply recognized that literature often reduces any thought to classifications, which follow from the mathematical concept of equivalence classes. Histograms are very immediate and informative wherever there is a phenomenon of interest which can yield measurements. This article should enable conclusions to be drawn about the equivalence of the underlying processes involved in sets of measurements for a very broad range of phenomena.

II. REFERENCES THROUGH SIMULATIONS

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

A. When the null hypothesis is true

III. REFERENCES 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.

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

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