Applied Mathematics, 2013, 4, 1028-1037
http://dx.doi.org/10.4236/am.2013.47140 Published Online July 2013 (http://www.scirp.org/journal/am)

Scientific Research

The Distribution of an Index of Dissimilarity for Two Samples from a Uniform Population

Giovanni Girone, Antonella Nannavecchia
Faculty of Economics, University of Bari, Bari, Italy
Email: nannavecchia.a@dss.uniba.it

Received October 11, 2012; revised January 6, 2013; accepted January 19, 2013

ABSTRACT

In this paper the authors study the sample behavior of the Gini's index of dissimilarity in the case of two samples of equal size drawn from the same uniform population. The paper present the analytical results obtained for the exact distribution of the index of dissimilarity for sample sizes $n \leq 8$. This result was obtained by expressing the index of dissimilarity as a linear combination of spacings of the pooled sample. The obtained results allow to achieve the exact expressions of the moments for any sample size and, therefore, to highlight the main features of the sampling distributions of the index of dissimilarity. The present study can enhance inferential statistical aspects about one of the main contributions of Gini.

Keywords: Index of Dissimilarity; Uniform Distribution; Spacings

1. Introduction

In 1915 Corrado Gini [1] introduced, as an index of dissimilarity between two groups of observations, the mean of the differences in absolute value between co-ranked observations. Fifty years later, in 1965, Gini [2] published a comprehensive and systematic overview of his own and of other authors results about dissimilarity.

We assume that two groups have equal number of observations n. Let $x_{11}, x_{12}, \cdots, x_{1i}, \cdots, x_{1n}$ and $x_{21}, x_{22}, \cdots,$ $x_{2i}, \cdots, x_{2n}$ be the set of the observations in the first and in the second group respectively. Let $x_{1(1)}, x_{1(2)}, \cdots,$ $x_{1(i)}, \cdots, x_{1(n)}$ and $x_{2(1)}, x_{2(2)}, \cdots, x_{2(i)}, \cdots, x_{2(n)}$ be the observations arranged in non-descending order of magnitude, respectively in the first and in the second group. Gini's simple index of dissimilarity is given by

$$D = \frac{\sum_{i=1}^{n} \left| x_{1(i)} - x_{2(i)} \right|}{n}.$$

If the two groups of observations are random samples, the distribution of the Gini index with samples drawn from the same population is of great interest for inferential purposes (e.g. to verify the homogeneity of two populations).

As far as we know, the distribution of the index of dissimilarity between two samples has not been studied in depth yet, maybe for the complex structure of the index which considers observations arranged in order of magnitude, co-ranked pairs and differences in absolute value. Instead the distribution of the index of dissimilarity between a sample and its population was obtained by Forcina and Galmacci [3] for the particular case of a discrete, equispaced and equidistributed population. The mean and variance of the distribution of the index of dissimilarity, both between two samples and between a sample and its population, have been obtained by Herzel [4] and Bertino [5] for some particular cases.

The purpose of this note is to study the sample behavior of the index of dissimilarity in the case of two samples from the same uniform population. In this paper we present the analytical results obtained for the exact distribution of the index of dissimilarity for sample sizes $n \leq$ 8 as well as for the exact general expressions of the main characteristic values (mean, variance, moments) of such a distribution for any sample size and their limiting values.

2. The Distribution of a Linear Combination of Spacings

Let $X_1, X_2, \cdots, X_i, \cdots, X_n$, be a random sample from a uniform population expressed by the probability density function (p.d.f.)

$$f(x)=1,\ 0<x<1, \qquad = 0,\ \text{elsewhere.} \tag{1}$$

The variables of the sample in nondescending order $X_{(1)},X_{(2)},\cdots,X_{(i)},\cdots,X_{(n)}$ determine $n+1$ spacings

$$X_{(1)}-0,X_{(2)}-X_{(1)},\cdots,X_{(i)}-X_{(i-1)},\cdots,$$
$$X_{(n)}-X_{(n-1)},1-X_{(n)}$$

denoted by $S_1,S_2,\cdots,S_i,\cdots,S_{n+1}$. These spacings are exchangeable random variables.

Let

$$Z=a_1S_1+a_2S_2+\cdots+a_iS_i+\cdots+a_{n+1}S_{n+1},$$

be their linear combination. Huffer and Lin [6] showed that the p.d.f. of Z is

$$h(z)=\sum_{i=1}^{p}\frac{1}{(r_i-1)!}\frac{\partial^{r_i-1}}{\partial y_i^{r_i-1}}\left(\frac{n(y_i-z)_+^{n-1}}{\prod_{j\neq i}(y_i-y_j)^{r_j}}\right), \tag{2}$$

in which $y_1,y_2,\cdots,y_j,\cdots,y_p$ are the distinct values of the weights a_i with frequencies $r_1,r_2,\cdots,r_j,\cdots,r_p$ while the sign +, appearing at the numerator of the ratio in the brackets, indicates that the function is nonzero only if it is positive. This result will be used for determining the p.d.f. of the index of dissimilarity.

3. The Index of Dissimilarity as a Linear Combination of Spacings

The case of two independent random samples from uniform population with p.d.f. given by expression (1), both of size n, can be brought to the case of one random sample of size $2n$ from the same uniform population. Let $X_{(1)},X_{(2)},\cdots,X_{(i)},\cdots,X_{(2n)}$ be the $2n$ variables of the pooled sample arranged in nondescending order of magnitude. They determine $2n+1$ spacings

$$X_{(1)}-0,X_{(2)}-X_{(1)},X_{(3)}-X_{(2)},\cdots,X_{(j)}-X_{(j-1)},\cdots,$$
$$X_{(2n)}-X_{(2n-1)},1-X_{(2n)},$$

which we denote by $S_1,S_2,\cdots,S_j,\cdots,S_{2n+1}$. Given all the $(2n,n)=(2n)!/(n!n!)$ subdivisions of the pooled sample into two samples of equal size n, let $h_1,h_2,\cdots,h_i,\cdots,h_n$ and $k_1,k_2,\cdots,k_i,\cdots,k_n$ be the ranks in the pooled sample of the variables in the first and in the second sample respectively. Moreover $p_i=\text{Min}(h_i,k_i)$ and $q_i=\text{Max}(h_i,k_i)$, for $i=1,2,\cdots,n$. It is easy to verify that the index of dissimilarity is given by

$$D=\frac{\sum_{i=1}^{n}\left(X_{(q_i)}-X_{(p_i)}\right)}{n},$$

or, when the differences are expressed in terms of spacings

$$X_{(q_i)}-X_{(p_i)}=\sum_{h=p_i+1}^{q_i}S_h,$$

by

$$D=\frac{\sum_{i=1}^{n}\sum_{h=p_i+1}^{q_i}S_h}{n},$$

clearly D is equal to the linear combination of spacings given by

$$D=\sum_{h=1}^{n+1}a_hS_h,$$

in which a_h is the relative frequency of the n intervals p_i+1-q_i that contain the hth spacing. The above procedure for each of the $(2n,n)$ subdivisions of the pooled sample should be used to determine the p.d.f. of the index of dissimilarity. To reduce the amount of computation it is appropriate to aggregate the subdivisions which bring to the same coefficients a_h. Due to the exchangeability of thespacings it is also appropriate to aggregate all subdivisions which bring to the same set offrequencies. In practice the $(2n,n)$ subdivisions are aggregated into groups characterized by the same set of frequencies. Denoting by $y_1,y_2,\cdots,y_j,\cdots,y_p$ the distinct values of coefficients a_h and by $r_1,r_2,\cdots,r_j,\cdots,r_p$ their frequencies (see Paragraph 2) we can apply formula (2) with the only variant of replacing n with $2n$. Taken together the subdivisions which have the same values (y_j,r_j) it is possible to proceed in an aggregate way by applying formula (2) only to homogeneous groups of subdivisions. These are in number of 2^{n-1} which allows to reduce considerably the amount of computation. Obviously, regard should be given to frequencies of the homogeneous subdivisions.

4. The Distributions of the Index of Dissimilarity for Sample Size Up to 8

To generate the sampling distribution of the index of dissimilarity two programs have been developed for *Mathematica software*. The first one generates, for each value of n, the subdivisions of the first $2n$ natural numbers into two subsets of size n and proceeds to identify homogeneous typologies and to define their frequencies. The second one starts, for each value of n, with such typologies and frequencies and gives, by applying formula (2), the p.d.f. of the index of dissimilarity. It should be said that heaviness of both calculation and resulting expressions led us to stop at $n\le 8$. The sampling densities of the index of dissimilarity (see Paragraph 2) are splines of order $2n-1$ with knots at points i/n for $i=1,2,\cdots,n-1$. They are also unimodal between 0 and 1, extreme values in which they assume value 0. The only exception is the degenerate case $n=1$.

For $n=1$

$$f(d)=2(1-d), \text{for } 0<d<1,$$
$$=0, \text{ elsewhere.}$$

For $n=2$

$$f(d)=\frac{16}{3}d\left(6-21d+19d^2\right), \text{ for } 0<d<\frac{1}{2},$$
$$=\frac{16}{3}(1-d)^3, \text{ for } \frac{1}{2}<d<1,$$
$$=0, \text{ elsewhere.}$$

For $n=3$

$$f(d)=\frac{324}{5}d^2\left(10-75d+195d^2-174d^3\right), \text{for } 0<d<\frac{1}{3},$$
$$=\frac{243}{20}(1-d)^5, \text{ for } \frac{1}{3}<d<\frac{2}{3},$$
$$=\frac{243}{20}(1-d)^5, \text{ for } \frac{2}{3}<d<1,$$
$$=0, \text{ elsewhere.}$$

For $n=4$

$$f(d)=\frac{8192d^3\left(1890-24570d+124362d^2-287931d^3+255625d^4\right)}{945}, \text{ for } 0<d<\frac{1}{4},$$
$$=\frac{8\left(-843+35196d-307440d^2+1283520d^3-3037440d^4+4193280d^5-3161088d^6+1008640d^7\right)}{945}, \text{for } \frac{1}{4}<d<\frac{1}{2},$$
$$=\frac{8\left(885-1092d-17136d^2+73920d^3-134400d^4+129024d^5-64512d^6+13312d^7\right)}{945}, \text{ for } \frac{1}{2}<d<\frac{3}{4},$$
$$=\frac{8192}{315}(1-d)^7, \text{ for } \frac{3}{4}<d<1,$$
$$=0, \text{ elsewhere.}$$

For $n=5$

$$f(d)=\frac{15625d^4\left(1161216-23224320d+192326400d^2-818424000d^3+1779880500d^4-1576012625d^5\right)}{36288},$$
$$\text{for } 0<d<\frac{1}{5},$$
$$=\frac{25\left(-184789+5314725d-54573300d^2+302410500d^3-1031073750d^4+2258943750d^5\right.}{108864}$$
$$-\frac{\left.3167062500d^6+2702812500d^7-1235390625d^8+212656250d^9\right)}{108864}, \text{ for } \frac{1}{5}<d<\frac{2}{5},$$
$$=\frac{25\left(208427-2703195d+17311500d^2-68533500d^3+178526250d^4-68533500d^3+178526250d^4\right.}{108864}$$
$$-\frac{\left.311456250d^5+360937500d^6 267187500d^7+114609375d^8-21718750d^9\right)}{108864}, \text{ for } \frac{2}{5}<d<\frac{3}{5},$$
$$=\frac{25\left(-27769+839745d-6308100d^2+23320500d^3-51108750d^4+71268750d^5-64312500d^6\right.}{108864}$$
$$+\frac{\left.36562500d^7+11953125d^8-11953125d^8+1718750d^9\right)}{108864}, \text{for } \frac{3}{5}<d<\frac{4}{5},$$
$$=\frac{1953125(1-t)^9}{36288}, \text{ for } \frac{4}{5}<d<1,$$
$$=0, \text{ elsewhere.}$$

For $n = 6$

$$f(d) = \frac{69984d^5 (2464000 - 70224000d + 860640000d^2 - 5771535000d^3}{9625}$$

$$+ \frac{22234481500d^4 - 46488497110d^5 + 41096463787d^6)}{9625}, \text{for } 0 < d < \frac{1}{6},$$

$$= \frac{(-39390245 - 596300430d + 56212110900d^2 - 985055920200d^3 - 985055920200d^3 + 8951408690400d^4}{2772000}$$

$$- \frac{51025364026560d^5 + 195684595666560d^6 - 516532350470400d^7 + 930778647955200d^8}{2772000}$$

$$- \frac{1098045852710400d^9 + 766505627458560d^{10} - 240526446437376d^{11})}{2772000}, \text{for } \frac{1}{6} < d < \frac{2}{6},$$

$$= \frac{(-268638245 + 8109363570t - 98322929100t^2 + 682896079800t^3 - 3088638669600t^4}{2772000}$$

$$+ \frac{9630608581440t^5 - 21202538797440t^6 + 33029606649600t^7 - 35735384044800t^8}{2772000}$$

$$+ \frac{25604654169600t^9 - 10945788733440t^{10} + 2116759578624t^{11})}{2772000}, \text{for } \frac{3}{6} < d < \frac{4}{6},$$

$$= \frac{(-15235805 + 275015730d - 2019609900d^2 + 8379142200d^3 - 22327034400d^4}{308000}$$

$$+ \frac{40613348160d^5 - 51839948160d^6 + 46651334400d^7 - 29099347200d^8}{308000}$$

$$+ \frac{12009254400d^9 - 2956124160d^{10} + 329204736d^{11})}{308000}, \text{for } \frac{4}{6} < d < \frac{5}{6},$$

$$= \frac{209952(1-t)^{11}}{1925}, \text{ for } \frac{5}{6} < d < 1,$$

$$= 0, \text{ elsewhere.}$$

For $n = 7$

$$f(d) = \frac{5764801d^6 (12809871360000 - 493180047360000d + 8375654327040000d^2}{100077120000}$$

$$- \frac{80940502843200000d^3 + 478835085704976000d^4 - 1728741146238943200d^5}{100077120000}$$

$$+ \frac{3517908542952383580d^6 - 3106403114212520549d^7)}{100077120000}, \text{for } 0 < d < \frac{1}{7},$$

$$= \frac{49(705190243035 - 48608926614885d + 1370336931078570d^2 - 21075409844719230d^3}{400308480000}$$

$$+ \frac{205665941648192025d^4 - 1375788745611094815d^5 + 6560947022693803740d^6}{400308480000}$$

$$- \frac{22672129451401209780d^7 + 56718764070636913245d^8 - 100876943897061006675d^9}{400308480000}$$

$$+ \frac{122406144806654198490d^{10} - 93386013407104353390d^{11}}{400308480000}$$

$$+ \frac{37532683619581030335d^{12} - 4652108914707165149d^{13})}{400308480000}, \text{for } \frac{1}{7} < d < \frac{2}{7},$$

$$= \frac{49\big(724389014235 - 26416076080485d + 495975655072170d^2 - 5590840507682430d^3}{400308480000}$$

$$+ \frac{41412935739200025d^4 - 213611632848829215d^5 + 794177103242462940d^6}{400308480000}$$

$$- \frac{2166089514477868980d^7 + 4351951762561047645d^8 - 6381666843892014675d^9}{400308480000}$$

$$+ \frac{6655537983365961690d^{10} - 4684154434862346990d^{11}}{400308480000}$$

$$+ \frac{1997194022429995935d^{12} - 390116369319886349d^{13}\big)}{400308480000}, \text{ for } \frac{2}{7} < d < \frac{3}{7},$$

$$= \frac{49\big(386280677245 - 7878437015995d + 77872173132390d^2 - 478219347415810d^3}{133436160000}$$

$$+ \frac{2005934860104175d^4 - 6039667172443905d^5 + 13424962922350980d^6}{133436160000}$$

$$- \frac{22336958123139660d^7 + 27870188207576715d^8 - 25798612790675725d^9}{133436160000}$$

$$+ \frac{17243777843522230d^{10} - 7888169349117330d^{11}}{133436160000}$$

$$+ \frac{2214635612061145d^{12} - 288256669921283d^{13}\big)}{133436160000}, \text{ for } \frac{3}{7} < d < \frac{4}{7},$$

$$= \frac{49\big(218508517245 - 4061620375995d + 37795598412390d^2 - 221061326295810d^3}{133436160000}$$

$$+ \frac{880868517704175d^4 - 2495708193883905d^5 + 5155725305710980d^6}{133436160000}$$

$$- \frac{7865792294019660d^7 + 8876783056856715d^8 - 7332802227475725d^9}{133436160000}$$

$$+ \frac{4317710449282230d^{10} - 1718909910957330d^{11}}{133436160000}$$

$$+ \frac{415268275931145d^{12} - 46034143903783d^{13}\big)}{133436160000}, \text{ for } \frac{4}{7} < d < \frac{5}{7},$$

$$= \frac{49\big(-22582874919 + 406770055419d - 3232969102818d^2 + 15243712005522d^3}{6671808000}$$

$$- \frac{47995298235885d^4 + 107152175090961d^5 - 175163892980076d^6}{6671808000}$$

$$+ \frac{212840606870892d^7 - 192597579807633d^8 + 128366236243245d^9}{6671808000}$$

$$- \frac{61318032201426d^{10} + 19895861688066d^{11}}{6671808000}$$

$$- \frac{3932902891827d^{12} + 357896140483d^{13}\big)}{6671808000}, \text{ for } \frac{5}{7} < d < \frac{6}{7},$$

$$= \frac{96889010407\,(1-t)^{13}}{444787200}, \text{ for } \frac{6}{7} < d < 1,$$

$$= 0, \text{ elsewhere.}$$

For $n=8$

$$f(d)=\frac{34359738368d^{7}\,(5028288890625-251414444531250d+5647327122937500d^{2}}{5028288890625}$$

$$-\frac{74128599182137500d^{3}+619867652797575000d^{4}-3372166777264740000d^{5}}{5028288890625}$$

$$+\frac{11629230369678054000d^{6}-23201017571436730200d^{7}+20470385435095442888d^{8})}{5028288890625},\ \text{for } 0<d<\frac{1}{8},$$

$$=\frac{(1785866560504985-98858130907162200d+1620233588120347200d^{2}}{20113155562500}$$

$$+\frac{10557759586230438400d^{3}-78774879801087375360 0d^{4}+142974105912210078105 60d^{5}}{20113155562500}$$

$$-\frac{1541670830134680813568 00d^{6}+1145320397304538123468800d^{7}}{20113155562500}$$

$$-\frac{6195682772044929721958400d^{8}+24964730331199205644697600d^{9}}{20113155562500}$$

$$-\frac{75264504448998097110958080d^{10}+168034884923532346864435200d^{11}}{20113155562500}$$

$$-\frac{270307440945096620716851200d^{12}+296825455668996496529817600d^{13}}{20113155562500}$$

$$-\frac{199459902900530267868364800d^{14}+61946662652365112780783616d^{15})}{20113155562500},\ \text{for } \frac{1}{8}<d<\frac{2}{8},$$

$$=\frac{(-10395733767175015+472434555633637800d-9801917876430052800d^{2}}{20113155562500}$$

$$+\frac{125944461384016038400d^{3}-1123561463695660953600d^{4}+7357990590660920770560d^{5}}{20113155562500}$$

$$-\frac{36458070001964567756800d^{6}+138927791089112894668800d^{7}}{20113155562500}$$

$$-\frac{410010935467157382758400d^{8}+936551720499415718297600d^{9}}{20113155562500}$$

$$-\frac{1641822440995120301998080d^{10}+2169256345186226877235200td^{11}}{20113155562500}$$

$$-\frac{2091386543607603475251200d^{12}+1389338588723279403417600d^{13}}{20113155562500}$$

$$-\frac{568841319984036367564800d^{14}+108244770034492549103616d^{15})}{20113155562500},\ \text{for } \frac{2}{8}<d<\frac{3}{8},$$

$$=\frac{(-12094942095795265+419329681293847800d-6550593682374772800d^{2}}{20113155562500}$$

$$+\frac{62257664756773158400d^{3}-405139576487996313600d^{4}+1917540769129588162560d^{5}}{20113155562500}$$

$$-\frac{6830701536285904076800d^{6}+18670785557317602508800d^{7}}{20113155562500}$$

$$-\frac{3951903669807313059840 0d^{8}+6482160311462537461760 0d^{9}}{20113155562500}$$

$$-\frac{81770292390085609390080d^{10}+77939848950584652595200d^{11}}{20113155562500}$$

$$-\frac{54354818157071512371200d^{12}+26190616878873418137600d^{13}}{20113155562500}$$

$$-\frac{7798288061683256524800d^{14}+1081819211351107567616d^{15}\Big)}{20113155562500},\ \text{for }\frac{3}{8}<d<\frac{4}{8},$$

$$=\frac{\Big(476110953668245-8404005584717400d+72266127037742400d^{2}}{6704385187500}$$

$$-\frac{368690549064947200d^{3}+1106957686081228800d^{4}-13695152708562124800d^{5}}{6704385187500}$$

$$-\frac{34864544499040256000d^{6}+22178074859510169600d^{7}-59187514072419532800d^{8}}{6704385187500}$$

$$+\frac{101932925346591539200d^{9}-123260133914821263360d^{10}+106598507655620198400d^{11}}{6704385187500}$$

$$-\frac{65019072070706790400d^{12}+26694919115715379200d^{13}}{6704385187500}$$

$$-\frac{6643496645138841600d^{14}+759154504863055872d^{15}\Big)}{6704385187500},\ \text{for }\frac{4}{8}<d<\frac{5}{8},$$

$$=\frac{\Big(138678600249637-3026020675659240d+31705558016274240d^{2}}{670438518750}$$

$$-\frac{206589742406494720d^{3}+925355068608122880d^{4}-3004552263085621248d^{5}}{670438518750}$$

$$+\frac{7298289851009597440d^{6}-13513030837248983040d^{7}}{670438518750}$$

$$+\frac{19250589909878046720d^{8}-21128554437756846080d^{9}}{670438518750}$$

$$+\frac{17742959702037233664d^{10}-11208493302452060160d^{11}}{670438518750}$$

$$+\frac{5161209629203496960d^{12}-1636889689514311680d^{13}}{670438518750}$$

$$+\frac{319966130020024320d^{14}-29078234263977984d^{15}\Big)}{670438518750},\ \text{for }\frac{5}{8}<d<\frac{6}{8},$$

$$=\frac{\Big(-12192935458067+233760762334680d-2043259180052160d^{2}}{26817540750}$$

$$+\frac{10869414530024960d^{3}-39517814160814080d^{4}+104311825731649536d^{5}}{26817540750}$$

$$-\frac{206943775415336960d^{6}+314693685576990720d^{7}-370262962360811520d^{8}}{26817540750}$$

$$+\frac{337377216755138560d^{9}-236297127330840576d^{10}+125003004515450880d^{11}}{26817540750}$$

$$-\frac{48370608882319360d^{12}+12930256742645760d^{13}}{26817540750}$$

$$\left.-\frac{2135801336954880d^{14}+164376988352512d^{15}}{26817540750}\right), \text{ for } \frac{6}{8}<d<\frac{7}{8},$$

$$=\frac{274877906944\left(1-t\right)^{15}}{638512875}, \text{ for } \frac{7}{8}<d<1.$$

The sampling distribution of the index of dissimilarity for $n=7$ is shown in **Figure 1**.

5. The Moments of the Sampling Distribution of the Index of Dissimilarity

On thee basis of the sampling distributions of the index of dissimilarity (see Paragraph 4), we obtained the values ofthe mean and of the second, third and fourth moment about the origin reported in **Table 1**.

Using the moments of the Dirichelet distribution, which is the model of the joint distribution of uniform spacings, as well as the expressions of the moments of a linear combination of equally distributed random variables, we obtained the mean and the second, third and fourth moment of the sampling distribution of the index of dissimilarity:

$$E\left(D\right)=\frac{2^{2n-1}}{\left(2n+1\right)\binom{2n}{n}},$$

$$E\left(D^2\right)=\frac{3+7n}{30n\left(n+1\right)},$$

$$E\left(D^3\right)=\frac{2^{2n}\left(21n^2+25n+2\right)}{64n^2\left(2n+1\right)\left(2n+3\right)\binom{2n}{n}},$$

$$E\left(D^4\right)=\frac{133n^3+294n^2+95n-18}{1260n^3\left(n+1\right)\left(n+2\right)}.$$

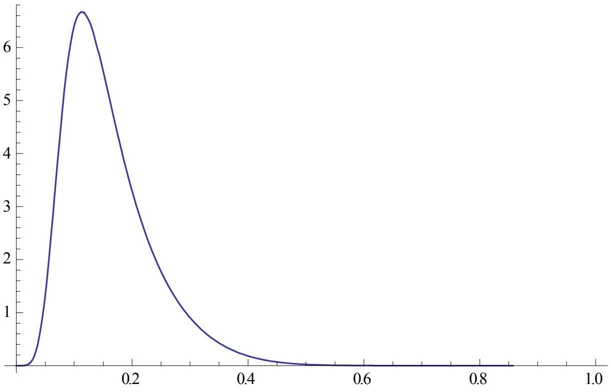

Figure 1. Sampling distribution of the index of dissimilarity for *n* = 7.

Based on these moments, we could study the main features of the sampling distribution of the index of dissimilarity starting from the mean. The trend of the mean with increasing sample size is shown in **Figure 2**.

It can be seen, the mean decreases for increasing n and it goes to zero for $n\to\infty$. The trend of the mean multiplied by $\sqrt{n}$ is increasing and converges to $\sqrt{\pi}/4$ (**Figure 3**).

The variance of the sampling distribution of the index of dissimilarity can be expressed as follows

$$\sigma_D^2=\frac{3+7n}{30n\left(n+1\right)}-\left[\frac{2^{2n-1}}{\left(2n+1\right)\binom{2n}{n}}\right]^2.$$

The trend of the standard deviation with increasing sample size is shown in **Figure 4**.

It can be seen that the standard deviation decreases with increasing n and it goes to zero for $n\to\infty$. The trend of the standard deviation multiplied by $\sqrt{n}$ is also

Table 1. Mean and moments of the sampling distribution of the index of dissimilarity for sample size up to 8.

Sample size	Moments about the origin			
	Mean	Second	Third	Fourth
1	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{10}$	$\frac{1}{15}$
2	$\frac{4}{15}$	$\frac{17}{180}$	$\frac{17}{420}$	$\frac{67}{3360}$
3	$\frac{8}{35}$	$\frac{1}{15}$	$\frac{19}{810}$	$\frac{271}{28350}$
4	$\frac{64}{315}$	$\frac{31}{600}$	$\frac{73}{4620}$	$\frac{2263}{403200}$
5	$\frac{128}{693}$	$\frac{19}{450}$	$\frac{2608}{225225}$	$\frac{1018}{275625}$
6	$\frac{512}{3003}$	$\frac{1}{28}$	$\frac{3632}{405405}$	$\frac{1661}{635040}$
7	$\frac{1024}{6435}$	$\frac{13}{420}$	$\frac{4288}{595595}$	$\frac{632}{324135}$
8	$\frac{16384}{109395}$	$\frac{59}{2160}$	$\frac{12368}{2078505}$	$\frac{2087}{1382400}$

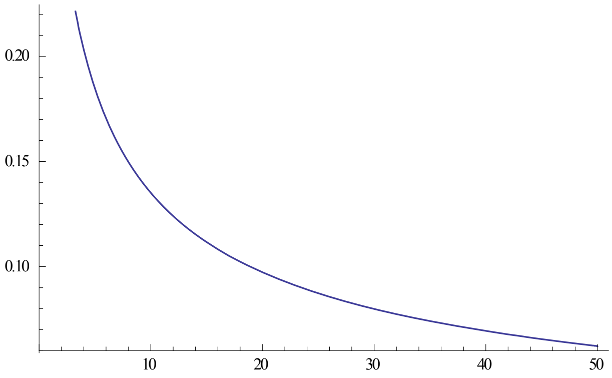

Figure 2. Trend of the mean with increasing sample size.

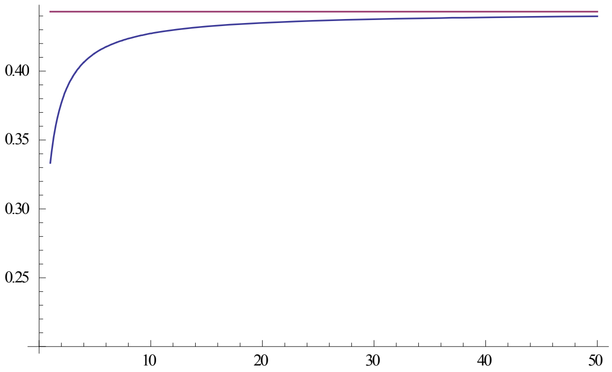

Figure 3. Trend of the mean multiplied by $\sqrt{n}$ with increasing sample size.

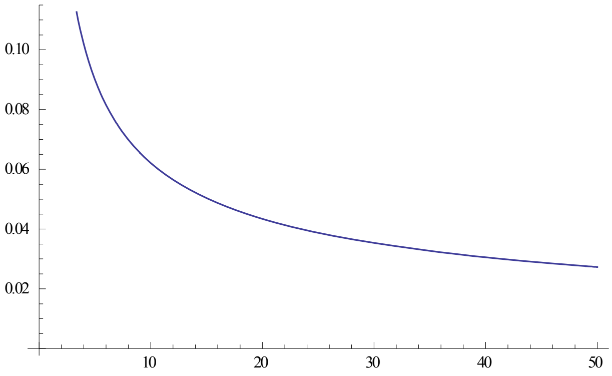

Figure 4. Trend of the standard deviation with increasing sample size.

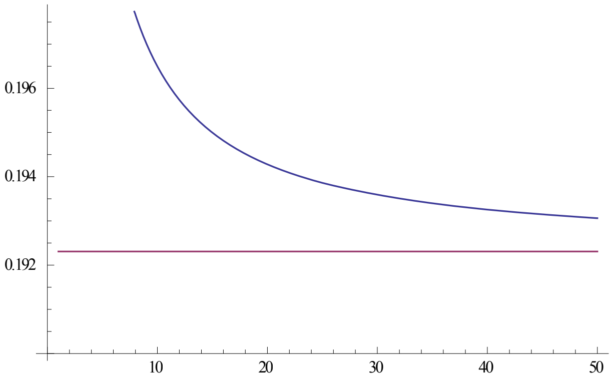

Figure 5. Trend of the standard deviation multiplied by $\sqrt{n}$ with increasing sample size.

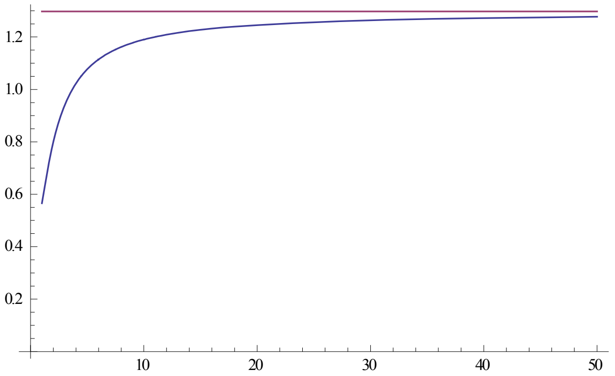

Figure 6. Trend of the skewness index with increasing sample size.

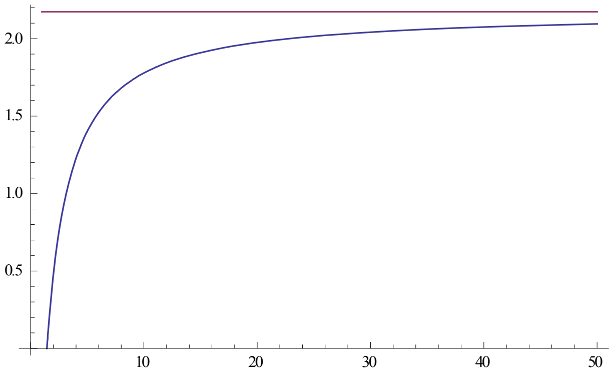

Figure 7. Trend of the excess of kurtosis index with increasing sample size.

decreasing and converges to $\sqrt{7/30-\pi/16}$ (**Figure 5**).

By converting moments about the origin into central moments it is possible to obtain indices of skewness and excess of kurtosis of the sampling distribution of the index of dissimilarity for a uniform population. Their trends with increasing sample size are shown in **Figures 6** and **7**.

It can be seen that distributions of the dissimilarity index are positively skewed and the skewness increases for increasing sample sizes up to the limit given by

$$\frac{3\sqrt{15\pi}\left(-119+40\pi\right)}{4\left(56-15\pi\right)^{3/2}}=1.297.$$

It is also clear that such distributions are leptokurtic and this leptokurtosis increases for increasing sample sizes up to the limit given by

$$\frac{-3328+45\left(119-30\pi\right)\pi}{\left(56-15\pi\right)^{2}}=2.174.$$

Table 2. Thresholds of the sampling distribution of the index of dissimilarity for sample size up to 8.

Sample Size	Levels of significance α			
	0.05	0.01	0.005	0.001
1	0.77639	0.90000	0.92929	0.96838
2	0.55995	0.70572	0.75254	0.83451
3	0.46038	0.58967	0.63236	0.71806
4	0.40115	0.51370	0.55458	0.64606
5	0.35941	0.45925	0.49438	0.55628
6	0.32987	0.42389	0.45876	0.52986
7	0.34104	0.39367	0.42633	0.49372
8	0.31588	0.36914	0.39995	0.46338

6. Thresholds of the Sampling Distributions of the Index of Dissimilarity

Thresholds at usual levels $\alpha = 0.05, 0.01, 0.005, 0.001$ are required to test statistical hypothesis that two samples come from the same uniform population. Thresholds have been calculated on the exact sampling distributions of the index of dissimilarity (see Paragraph 4) and their values are reported in **Table 2**. For larger sample sizes can be used the approximate conservative thresholds

$$\alpha = 0.05 \quad 0.100\sqrt{n},$$

$$\alpha = 0.01 \quad 0.116\sqrt{n},$$

$$\alpha = 0.001 \quad 0.126\sqrt{n},$$

$$\alpha = 0.005 \quad 0.147\sqrt{n}.$$

7. Conclusions

In this note we obtained the distributions of the index of dissimilarity in the case of two random samples of small size from a uniform population. This result was obtained by expressing the index of dissimilarity as a linear combination of spacings of the pooled sample. The exact expressions, in the form of splines of order $2n - 1$, were obtained for samples of size ≤ 8. Although it would be possible to go further we stopped at this size because of the heaviness in processing and in the resulting expressions. This limit could be overcome by using some already set calculus programs. The obtained results allow to achieve the exact expressions of the moments and, therefore, to highlight the main features of the sampling distributions of the index of dissimilarity.

Beyond the problem of dealing with larger sample sizes, open problems are also those considering the case of two samples with different sizes as well as the case of other distribution models for population.

We believe that the present work has re-opened a research strand able to enhance also the inferential statistical aspects about one of the fundamental contributions of Gini.

REFERENCES

[1] C. Gini, "Di Una Misura Della Dissomiglianza tra Due Gruppi di Quantità e Delle Sue Applicazioni Allo Studio Delle Relazioni Statistiche," *Proceedings of the R. Venetian Institute of Sciences*, *Literatures and Arts*, Vol. 74, 1914, pp. 185-213.

[2] C. Gini, "La Dissomiglianza," *Metron*, Vol. 24, No. 1-4, 1965, pp. 85-215.

[3] A. Forcina and G. Galmacci, "Sulla Distribuzione Dell' Indice Semplice di Dissomiglianza," *Metron*, Vol. 32, No. 1-4, 1974, pp. 361-378.

[4] A. Herzel, "Il Valor Medio e la Varianza Dell'Indice Semplice di Dissomiglianza Negli Universi dei Campioni Bernoulliano ed Esaustivo," *Library of Metron*, *Series C*, *Notes and Reports*, Vol. 2, 1963, pp. 199-224.

[5] S. Bertino, "Sulla Media e la Varianza Dell'Indice Semplice di Dissomiglianza Nel Caso di Campioni Provenienti da una Stessa Variabile Casuale Assolutamente Continua," *Metron*, Vol. 30, No. 1-4, 1972, pp. 256-281.

[6] W. F. Huffer and C. T. Lin, "Spacings, Linear Combinations of," In: *Encyclopedia of Statistical Sciences*, John Wiley & Sons Inc., New York, 2006, pp. 7866-7875. doi:10.1002/0471667196.ess5049.pub2