Statistical Tests Based on Gini Index and Gini Mean Difference

Magdalena Niewiadomska-Bugaj
Department of Statisics, Western Michigan University, USA
m.bugaj@wmich.edu
Teresa Kowalczyk
tkow@ipipan.waw.pl
Institute of Computer Science, Polish Academy of Sciences, Poland
Hend Auda
Department of Statistics, Western Michigan University, USA

February 7, 2005

Abstract

Our paper will be devoted to tests of hypotheses based on the Gini index or Gini mean difference. Being a counterpart of a coefficient of variation (σ/μ), Gini index can be used as a statistic in nonparametric tests such as goodness of fit test (Jammalamadaka and Goria, 2004), two sample test (Niewiadomska-Bugaj, 2003), and proposed test for independence. These tests use as a test statistic Gini obtained for random spacings, rank spacings and for rank differences respectively. Properties and performance of these tests will be presented, and possible further extensions will be discussed.

1 Introduction.

Gini mean difference (GMD) and Gini index (GI) have been proposed as measures of income inequality almost 100 years ago. Since 1921 when the Gini's paper "Measurement of Inequality of Incomes" was published, they became well established measures of inequality, widely used in economical research. There exists voluminous literature on the subject of Gini index regarding computational aspects, properties, and especially decomposition which enables studies of the inequality in subpopulations and making various comparisons. In our paper we want to focus on the statistical tests that are based on the Gini index (or Gini mean difference), in a sense that they use measure of inequality for assessing deviation from the "standard" or "ideal" specified by the null hypothesis. Such methods not only can be more powerful than several existing ones, but they often have other advantages. For example, Gini mean difference

as a measure of variation does not require any information about the "center" measured usually by either a mean or a median. Therefore methods based on Gini mean difference are often rotationally invariant and as such applicable to directional data.

2 Two-Sample Nonparametric Test

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, be two random samples from continuous distributions with cdf's F_X and F_Y respectively. Moreover, let R_i^{XY} denote the rank of $X_{i:m}$ (i-th order statistics in $X_1, \ldots, X_m$) in the combined sample. Then $S_i = S_i^1$ for $i = 1, \ldots, m+1$, so called simple rank spacings, are numbers of Y observations falling between two consecutive order statistics $X_{i-1:m}$ and $X_{i:m}$, and can be obtained as $S_i^1 = R_i^{XY} - R_{i-1}^{XY} - 1$ for $i = 1, \ldots, m$, where $R_0^{XY} = 0$, and $S_{m+1}^1 = m + n - R_m^{XY}$. Similarly, rank spacings of order k ($m+1$ has to be divisible by k) are defined as $S_i^k = R_{ki}^{XY} - R_{k(i-1)}^{XY} - k, i = 1, \ldots, t-1$, and $S_{t+1} = m+n-(k-1)-R_{(t-1)k}$where $t = (m+1)/k$. Rank spacings of order k are numbers of observations among $Y_1, \ldots, Y_n$ falling between every k-th order statistic $X_{1:m}, \ldots, X_{m:m}$.

Among all tests of the form

$$T_{m,n} = \sum_{i=m}^{n} h_N(S_i),$$

symmetric in rank spacings S_i, Holst and Rao (1981) found test based on $\sum_{i=1}^{m}(S_i)^2$ proposed by Dixon (1940) to be asymptotically most powerful. Similarly, among tests symmetric in rank spacings S_i^k the test $\sum_{i=1}^{m}(S_i^k)^2$is asymptotically locally most powerful (see Rao and Schweitzer (1982)). However, these tests were not most powerful when compared with other tests based on ranks. Kaigh (1994) proposed a two-sample test based on rank spacings which was shown to be, for most of the alternatives, more powerful than all widely used other two-sample rank tests such as Anderson-Darling, Cramer von Mises, Kolmogorov-Smirnov, Kruskall-Wallis and Mood tests.

Following approach of Boos (1986), who proposed an omnibus two-sample rank test based on squared rank components, Kaigh (1994) proposed an omnibus test based on squared rank spacings components. Test statistic

$$\Psi_s^{XY2} = \sum_{1 \le s \le p} ([Z_{s;N}^X]^2 + [Z_{s;N}^Y]^2)/2,$$

where

$$Z_{s;N}^X = -\sqrt{\frac{(m+1)(m+2)}{n(N+1)}} \sum_{1 \le i \le m+1} \pi_{s,m}(i) S_i^X,$$

and $Z_{s;N}^Y$ is defined analogously, is a linear combination of simple ($k = 1$) rank spacings with Hahn polynomial vector weight functions $\pi_{s,m}(i)$ obtained

from orthogonal decomposition of a Dixon statistic. First four components ($s = 1, 2, 3, 4$) can be interpreted as nonparametric measures of location, scale, skewness and kurtosis, respectively. Kaigh considered two aggregate statistics: Ψ_2^{XY2} with two first components and Ψ_4^{XY2} with four first components since inclusion of all rank spacings components apparently dilutes the power of aggregate statistics (see Kaigh (1994), also for comparison of Ψ_2^{XY2}, Ψ_4^{XY2} and other two-sample rank tests).

Discrepancy between two distributions can be assessed by a Gini index. Kowalczyk (1994) has shown that the problem of divergence of any two mutually absolutely continuous distributions F_Y and F_X is equivalent to the problem of inequality of the distribution of $h(X)$ where $h = dF_Y/dF_X$. If distributions F_X and F_Y are identical then $h \equiv 1$. The inequality of the distribution of $h(X)$ measures its departure from degenerate distribution of $h(X)$, where $P(h(X) = 1) = 1$. When $h = dF_Y/dF_X \equiv 1$ we have $P(h(X) = 1) = 1$, and consequently $GI(h(X)) = 0$. Otherwise, $P(h(X) = 1) < 1$ which corresponds to $GI(h(X)) > 0$. The more different are distributions F_X and F_Y, the more inequality is in the distribution of $h(X)$. Relative frequencies S_i^1/n provide histogram-like density estimator for $h(X)$, and variables $S_1^1, \ldots, S_{m+1}^1$ have multinomial distribution with $p_i = \frac{1}{m+1}$ for $i = 1, \ldots, m+1$. Gini index, applied to rank spacings $S_1^1, \ldots, S_n^1$, equals

$$\begin{aligned} gi_n &= \sum_j \sum_i \frac{|S_i^1 - S_j^1|}{2mn} = \frac{\sum \sum_{i>j} (S_{i:m+1}^1 - S_{j:m+1}^1)}{mn} \\ &= \frac{2\sum_i i S_{i:m+1}^1 - n(m+2)}{mn}. \end{aligned}$$

When $X_{m:m} < Y_{1:n}$ ($Y_{n:n} < X_{1:m}$), so samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are completely separated then $S_1^1 = \cdots = S_m^1 = 0, S_{m+1}^1 = n$, and

$$gi_n = \frac{2n(m+1) - n(m+2)}{mn} = 1.$$

When $m \leq n$ and $Y_{1:n} < X_{1:n-1} < Y_{2:n} < X_{2:n-1} < \cdots < X_{m:m} < Y_{m:n}$ then $S_1^1 = \ldots = S_m^1$ and $S_{m+1}^1 = n - m$ and

$$gi_n = \frac{2\sum_{i=1}^m i + 2(m+1)(n-m) - n(m+2)}{mn} = \frac{n - (m+1)}{n},$$

which equals 0 when $n = m + 1$. Similarly, CI for spacings of order k can be obtained as

$$gi_n^k = \frac{\sum_i \sum_j |S_i^k - S_j^k|}{2n(t-1)} = \frac{2\sum_{i=1}^t S_{i:t}^k - n(t+1)}{n(t-1)}.$$

The test we are proposing, is based on Gini index, a non-symmetric function of rank spacings. We will compare its power with that of a Dixon test (as the most powerful among symmetric functions of rank spacings) for different values

Table 1: Power comparison for selected two-sample Kaigh test and Dixon D^k and Gini G^k tests for order or rank spacings $k = 1, 2, 4, 5, 10$ for $m = 19, n = 20$

Test	B(2, 3)	B(2, 4)	B(2,9)	0.5[B(2, 19) +B(19, 2)]	0.5[B(4, 17) +B(19, 2)]
$k = 1$					
Kaigh	0.402	0.663	0.988	0.711	0.332
Gini	0.230	0.350	0.796	0.667	0.523
Dixon	0.206	0.294	0.715	0.576	0.441
$k = 2$					
Gini	0.249	0.390	0.880	0.734	0.547
Dixon	0.249	0.371	0.833	0.675	0.503
$k = 4$					
Gini	0.327	0.508	0.935	0.779	0.538
Dixon	0.303	0.463	0.911	0.723	0.490
$k = 5$					
Gini	0.297	0.467	0.938	0.712	0.438
Dixon	0.293	0.459	0.932	0.689	0.417
$k = 10$					
Gini	0.239	0.473	0.962	0.003	0.004
Dixon	0.239	0.473	0.962	0.003	0.004

of k, and with Kaigh test Ψ_2^{XY2}. Results of the simulation study are summarized in Table 1.

We were especially interested in tests' performance in the case of deviation from symmetry and change in modality. Distribution of one population was uniform U[0, 1]=B(1, 1) while as other distribution we used B(2, 3), B(2, 4), B(2, 9), and two bimodal distributions - mixtures of skewed beta distributions: $0.5[B(2, 19) + B(19, 2)]$ and $0.5[B(4, 19) + B(19, 2)]$, where the last distribution was additionally not symmetric. Sample were of sizes $m = 19$ and $n = 20$, generated 5000 times.

As expected (see Rao and Seturaman (1975)) power of Gini and Dixon tests firstly increases with the order of rank spacings k, attains its maximum for the moderate k and then decreases. Gini test seems to be more powerful than the Dixon test regardless the order of rank spacings $k < 10$. While the Kaigh test is most powerful of the three tests being compared when $k = 1$, Gini and Dixon tests applied to rank spacing of order $k > 1$ have a higher power. When $k = 10$, information is reduced heavily and both tests are no longer of a practical use in the case of change in modality - notice a dramatic drop in power which becomes even less than 0.01.

3 Independence Test

Let us now consider a sample of size n from continuous bivariate distribution $(X_1, Y_1), \ldots .(X_n, Y_n)$ with $X_{Y_{i:n}}$ denoting X observation that corresponds to i-th in magnitude Y observation $(Y_{i:n})$. After ranking X-values and Y-values separately, we obtain n pairs of ranks $(R_1, S_1), \ldots, (R_n, S_n)$ and n their differences $D_1, \ldots, D_n$ where $D_i = |R_i - S_i|$. Small variation among D_i's indicates positive association, large variation corresponds to negative association while moderate values indicate lack of association often referred to as independence. These small, large, and moderate dispersions of absolute rank differences correspond to small, large and moderate values of a Gini mean difference respectively.

Small variations in D_i's usually mean that they are equal or close to zero what is the case in positive association. One can consider, at least theoretically, the case when all absolute rank differences are equal to a positive value. Gini mean difference would in such case also be equal zero, and such case corresponds to positive association in the presence of a covariate, so small value of a statistic is justified. To evaluate the performance of the nonparametric Gini test of independence (association) based on statistic

$$\begin{aligned} \gamma_n &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|D_i - D_j|}{2n(n-1)} = \frac{\sum\sum_{i>j}(D_{i:n} - D_{j:n})}{n(n-1)} \\ &= \frac{2\sum_{i=1}^{n} iD_{i:n} - n(n+1)}{n(n-1)} \end{aligned}$$

we performed a preliminary simulation study that compared its power with that of Spearman (1904) and Hoeffding (1948) tests - widely used powerful nonparametric procedures, test based on a Gini coefficient of association

$$C_n = \frac{1}{M_n}\sum_{i=1}^{n}||R_i + S_i - (n+1)| - |R_i - S_i||, \tag{1}$$

where $M_n = n^2/2$ or $M_n = (n^2-1)/2$ depending on whether n is even or odd, and a test with a statistic

$$G(X,Y) = \frac{\sum_{i=1}^{n}(2i-1-n)X_{Y_{i:n}}}{\sum_{i=1}^{n}(2i-1-n)X_{i:n}} \tag{2}$$

being a sample counterpart and consistent estimator of

$$\Gamma(X,Y) = \frac{Cov(X, F_Y(Y))}{Cov(X, F_X(X)}.$$

Statistic (1) which was introduced by Gini (1914), then studied by Cifarelli and Regazzini (1977) and by Nelsen (1998). Coefficient (2) was proposed by Schechtman and Yitzhaki (1987) as a measure of association and is also based on a Gini mean difference.

The results of our simulation study are summarized in Table 2. Power was obtained based on 5000 repetitions, sample size $n = 20$. For alternative distributions we used Frank's family of bivariate distributions with the following joint cdf $H_\alpha(x, y)$:

$$H_\alpha(x, y) = \log_\alpha \left(1 + \frac{(\alpha^x - 1)(\alpha^y - 1)}{\alpha - 1}\right)$$

(see Frank (1979)).

Table 2: Power comparison for selected tests of independence (association). Alternative distributions were members of Frank's family corresponding to $a = 1.5, 3, 6$, and 9, where $\alpha = \exp(-a)$.

Test	$a = 1.5$	$a = 3$	$a = 6$	$a = 9$
Gini (new)	0.254	0.611	0.973	0.999
Spearman	0.258	0.613	0.972	0.999
Hoeffding	0.286	0.600	0.954	0.996
Gini association	0.088	0.152	0.463	0.790
Gini correlation	0.196	0.297	0.402	0.458

Results in Table 2 indicate that the power of Gini and Spearman tests is very similar and while it is slightly less than power of the Hoefding test in the case of a relatively weak association ($a = 1.5$), the trend reverses when the strenght of association increases. Also, it is clear that other two tests (using as a test statistic Gini coefficeint of association and Gini correlation) are consistently less powerful.

Frank's family of distributions is a family with a positive likelihood ratio dependence and as such provides strong form of association. In further studies other dependence structures should be considered as well.

4 Goodness-of-Fit Test

We want to mention here goodness-of-fit test introduced by Jammalamadaka and Goria (2004). They proposed to apply Gini index to spacings and obtained test statistic

$$G_n = \sum_{i=1}^{n+1}\sum_{j=1}^{n+1} |D_i - D_j| = 2\sum_{i=1}^{n}[\sum_{j=i+1}^{n+1} (D_{j:n} - (n - i + 1)D_{i:n})] \quad (3)$$

$$= 4\sum_{i=1}^{n} iD_{i:n} - 2(n + 2), \quad (4)$$

where $D_i = X_{(i+1):n} - X_{i:n}, i + 1, \ldots, n$ and $D_{n+1} = 1 - X_{n:n}$ are obtained from the original data after they were transformed by the cdf of the continuous distribution specified by the null hypothesis. Jammalamadaka and Goria

(2004) provide exact and asymptotic distribution for the test statistic and study its performance. They also present results of a power study where they compare Gini test with three other competitors. Also, Gail and Gastwirth (1978) proposed scale free test for exponentiality based on a Gini index.

5 Final Remarks

In our paper we have shown three statistical tests of different hypotheses all based on the Gini index or the Gini mean difference. Other tests based on the Gini index can be developed as well. For example, test for the symmetry, or a test that compares variability (dispersion) in two or more populations that would be especially applicable in circular data could be investigated. Gini proposed his index to assess dispersion independently of any measure of center or location. Understood as a mean absolute difference Gini index is especially attractive for assessment of variation in circular data, since it is rotationally invariant.

6 References:

Boos, D.D. (1986). Comparing K Populations With Linear Rank Statistics. *Jour. Amer. Statist. Association*. **81**, 1018 - 1025.

Cifarelli, D.M., and Regazzini, E. (1977). On a Distribution-Free Test of Independence Based on Gini's Rank Correlation Coefficient. In: Barra, J.R. et al., eds. *Recent Developments in Statistics*. Amsterdam: North Holland, 375-385.

Dixon, W.J. (1940). A Criterion for Testing the Hypothesis That Two Samples Are From the Same Population. *The Annals of Mathematical Statistics*.**11**, 199-204.

Frank, M.J. (1979). On the Simultaneous Associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Math.* **19**, 194-226.

Gail, M.H., and Gastwirth, J.L. (1978). A Scale-free Goodness-of-fit Test for the Exponential Distribution based on the Gini Statistic. *J.R. Statist. Soc.* bf B, bf 40, 350-357.

Gini, C. (1914). *L'Ammontare e la Composizione della Ricchezza delle Nazioni*, Bocca, Torino.

Gini, C. (1921). Measurement of Inequality of Incomes. *The Economic Journal* **31**, 124-126.

Hoeffding, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19**, 546-557.

Holst, L., and Rao, J.S. (1981). Asymptotic Spacings Theory With Applications to the Two Sample Problem. *Canadian Jour. of Statistics.* **9**, 79-89.

Jammalamadaka, S.R., and Goria, M.N. (2004). A Test of Goodness of Fit Based on Gini's Index of Spacings. *Statistics and Probability Letters.* **68**, 177-187.

Kaigh, W.D. (1994). Distribution-Free Two-Sample Tests Based on Rank Spacings. *Jour. Amer. Statist. Association.* **89**, 159-167.

Kowalczyk, T. (1994). A unified Lorenz-type approach to divergence and dependence. *Dissertationes Mathematicae.* CCCXXXV.

Nelsen, R. (1998). Concordance and Gini's Measure of Association. *Nonparametric Statistics.* **9**, 227-238.

Niewiadomska-Bugaj, M. (2003). Statistical Analysis Related to Gini Mean Difference. In:*Proceedings of the First Brazilian Conference on Statistical Modelling in Insurance and Finance*, University of Sao Paulo, Brazil. 553 - 569.

Rao, J.S. and Sethuraman, J. (1975). Weak convergence of empirical distribution functions of random variables subject to perturbation and scale factors. *Ann. Statist.* **3**, 299-313.

Rao, J.S. and Schweitzer, R.L. (1982). On Tests for the Two-Sample Problem Based on Higher Order Spacing-Frequencies. In: Matusita, K. ed. *Statistical Theory and Data Analysis.* North-Holland Publishing Co., 583-618.

Spearman, C. (1904). The proof and measurement of Association between two things. *Amer. J. Psychol.* **15**, 72-107.

Schechtman, E., and Yitzhaki, S. (1987). A Measure of Association Based on Gini's Mean Difference. *Commun. Statist. Theor. Meth.*, 207-231.