The Greatest Deviation Correlation Coefficient and its Geometrical Interpretation

By Rudy A. Gideon

The University of Montana

The Greatest Deviation Correlation Coefficient (GDCC) was introduced by Gideon and Hollister (1987). The GDCC was shown to possess all the properties of a rank based correlation coefficient, a new tied value procedure was introduced, the null distribution for independence was given, and a population interpretation was given. This paper expands the exact distribution up to sample sizes 15, gives a more intuitive definition based on the graph of the data, and demonstrates the geometrical uniqueness of the definition by 90^0 rotations of the graph. An analogous graph is used to give a geometric definition of the population or theoretical value of GDCC. Four correlation coefficients, GDCC, Kendall's, Spearman's, and an absolute value rank correlation coefficient are computed on the scatterplot of the ranked data by simple counting methods. The significance of GDCC as a robust correlation coefficient is illustrated and the use of the asymptotic distribution is demonstrated.

1. The Mathematical Definition of GDCC

This paper brings together components of the GDCC in order to make it a useful tool for studying the relationships between variables. We start with the basic or mathematical definition.

The method will be illustrated with the YMCA data that appeared in the 1987 article. The data came as the ranks (won-lost record versus Sportsmanship) of 16 teams of 4^{th} and 5^{th} graders in a YMCA basketball league. The Sportsmanship rank was a summary tally of a weekly evaluation by team coaches and officials. The data is reproduced here in Table 1.

TABLE 1: YMCA team ranking in Sportsmanship and won-lost record

p_i	14	11	16	2	12	13	7	9	10	3	8	1	15	6	4	5
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

The Won-Lost rank data is the lower row and can be considered the independent variable while the upper row is the Sportsmanship rank and is taken as the dependent variable. By viewing the graph of this data (see Graph) one immediately notices that (4,2) and (13,15) are unusual data points. We write the data in the form $\{i, p_i\}, i = 1,2,3,...,16$; that is, numerically ordered by the won-lost rankings.

The general definition of the GDCC, with sample size n, is

$$r_{gd} = \frac{\max\limits_{1\le i\le n} \sum\limits_{j=1}^{i} I(n+1-p_j > i) - \max\limits_{1\le i\le n} \sum\limits_{j=1}^{i} I(p_j > i)}{\left[\frac{n}{2}\right]} \qquad (1)$$

where the denominator is the greatest integer function evaluated at n/2, and *I* is the indicator function. The indicator function is 1 is the expression is true and 0 if false. For small sample sizes this is easy to compute by hand, and the layout is shown for the YMCA data in Table 2. Three columns are necessary, $\{i, p_i, n+1-p_i\}$ where the first column of the original data must be in numerical order. From these columns the elements in the summations in the formula above are computed.

This computational operation follows:

Column 1 consists of values of i =1,2,3…16 and column 2 is the corresponding values of p_i (Sportsmanship rank). In order to compute $\sum_{j=1}^{i} I(p_j > i)$, labeled I-sum-R in column 3, a simple procedure is used. For row 1 (i =1) simply count how many p-ranks (the p column) there are in rows 1 and above that are greater than i=1. Since p_1 =14, there is only one, namely 14 itself, so the 3rd column entry is 1. For row 2 (i =2), p_2 =11 and both 14 and 11 (p_1 and p_2) exceed i =2 and hence the corresponding column 3 value is 2. Continue down the column. One last example: for row 7 (i =7) we examine the number of entries from $\{p_1,p_2, \ldots ,p_7\}$ = {14,11,16,12,13} which exceed 7. There are five. In the exact same manner, columns 1 and 4 (the reverse column) are used to calculate the values in column 5 whose formula is the left hand side of formula (1). The maximums of columns 3 and 5, 6 and 3, are conveniently put on the bottom of columns 3 and 5. Then

$$r_{gd} = \frac{\max(I-sum-L) - \max(I-sum-R)}{\left[\frac{16}{2}\right]} = \frac{3-6}{8} = \frac{-3}{8}.$$

TABLE 2: Calculation of the Greatest Deviation Correlation Coefficient

column 1	column 2	column 3		column 4	column 5
i	p_i	I-sum-R		n+1-p_i	I-sum-L
1	14	1		3	1
2	11	2		6	2
3	16	3		1	1
4	2	3		15	2
5	12	4		5	2
6	13	5		4	1
7	7	5		10	2
8	9	6		8	2
9	10	6		7	2
10	3	5		14	2
11	8	4		9	2
12	1	3		16	3
13	15	3		2	3
14	6	2		11	2
15	4	1		13	1
16	5	0		12	0
maximum		6			3

2. The Geometrical Definition of GDCC

We now make the transition from the formula calculation of r_{gd} to a graphical counting technique. The scatterplot of the YMCA data $\{i, p_i\}, i = 1,2,3,...,16$ is given in figure 1. The lines with slopes ± 1 have been added to facilitate the counting. First, the partial sum $\sum_{j=1}^{i} I(p_j > i)$ is merely the count among the first i p_i-ranks of those that are strictly greater than i; i.e., the data points in the scatterplot strictly above and to the left of the point (i,i) on the +1 diagonal. Points on the right hand side of the defining rectangle are included in the count; thus, the counting region is closed on the right but open on the bottom. A maximum of 6 is achieved at i=9, and the borders of this rectangular region are shown on the graph. Note that the point (8,9) on the lower is not counted because of the strict inequality whereas the point (9,10) is counted. Thus, there are 6 points for the maximum. If (8,8) is used to define a counting region, the maximum of 6 is also achieved.

Second, the partial sum $\sum_{j=1}^{i} I(n+1-p_j > i) = \sum_{j=1}^{i} I(p_j < n+1-i)$ is the count among the first i p_i-ranks of those that are strictly less than n+1-i = 17-i; i.e., those points that are on the vertical or to the left and strictly below the point (i, 17-i) on the –1 diagonal. A maximum of 3 is achieved at i = 12 with the region containing the 3 points shown in the figure. The maximum is also achieved at *i*=13, as is easily seen in the figure.

It is clear that the two parts of the numerator of r_{gd} can be calculated by traversing the ± 1 diagonals with moving rectangles and determining which rectangular areas contain the maximal number of data points. The number of points in each rectangular region $i = 1,2,3,...16$ will correspond exactly to the numbers in the I-sum-R and I-sum-L columns of Table 2.

Geometrically, it may appear that some information is lost or a different value of r_{gd} could be obtained by traversing the ± 1 diagonals on the other sides. The next section relates this question to the general properties of r_{gd}. The geometrical feature of this section was used in obtaining the asymptotic distribution of r_{gd}. Reference=?

3. Geometrical Uniquenes of r_{gd}

The information in the scatterplot should remain the same if the graph is rotated 90^0 and r_{gd} is calculated over on this new orientation. When the graph is rotated 90^0, let the axes be relabeled in the usual left to right and vertical manner and recompute r_{gd}.

This 90^0 rotation will be done three consecutive times. A fourth rotation brings back the original graph.

Let the notation $(\underline{-1},\overline{+1})$ denote the relationship of the counting rectangles to the original orientation of the axes. Thus, $(\underline{-1},\overline{+1})$ means lower rectangle for the –1 diagonal and upper rectangle for +1 diagonal, the scheme in figure 1. We now explain what happens

to the data under three consecutive 90^0 rotations; the orientation of the counting rectangles relative to the original figure, and the new value of r_{gd}. In Table 3 the first row is the data after a rotation, the row the orientation of the counting rectangles relative to the original graph, the third row is the value of r_{gd}, and the fourth row is what is effectively being calculated as related to the original data, x = i, won lost ranks, and y = p_i, the ranks of sportsmanship.

Table 3:Rotational Effects on Figure 1

	original	first rotation	second rotation	third rotation
data	$\{i,p_i\}$	$\{17\text{-}p_i,i\}$	$\{17\text{-}i,17\text{-}p_i\}$	$\{p_i,17\text{-}i\}$
diag sides	$(\underline{-1},\overline{+1})$	$(\overline{-1},\overline{+1})$	$(\overline{-1},\underline{+1})$	$(\underline{-1},\underline{+1})$
r_{gd} value	-3/8	3/8	-3/8	3/8
$r_{gd}(.,.)$	(x,y)	(x,-y)	(-x,-y)	(-x,y)

Not only does r_{gd} maintain the same absolute value, but the counting rectangles while traversing the ± 1 diagonals give exactly the same sequences of numbers. Thus, the I-sum-R and I-sum-L columns of the rotations are exactly the same as for the original data. However, they may be in reverse order. This connects the geometry to property 4 of Section 3 and fact (d) on page 654 of the [1987] paper. The reader is invited to perform these calculations by rotating the figure, relabeling the axes and recomputing r_{gd}.

4. Population Definition of GDCC

This section repeats the population interpretation of r_{gd} in the 1987 paper and interprets it relative to the geometrical ideas of the sample definition. Let (X,Y) be a continuous bivariate random variable with marginal distribution functions F and G, respectively. Let $(U,V) = (F(X),G(Y))$, and define the Copula function, Nelson (1999),

$$P(U \le t, V < 1-t) = C(t,1-t)$$
$$P(U \le t, V > t) = C(t,1) - C(t,t).$$

A graph of the density of (U,V) is given in which the bivariate normal from which it came has means 0 and standard deviations of $\sqrt{2}$ and $\sqrt{32}$, and covariance 3. The population value of r_{gd} is

$$r_{gd} = 2\sup_{0<t<1} C(t,1-t) - 2\sup_{0<t<1}[C(t,1) - C(t,t)]. \qquad (2)$$

The right hand side of the population value of r_{gd} corresponds to the right hand side of the statistic r_{gd}, $\max \sum_{j=1}^{i} I(p_j > i)$. The left hand side $\sup C(t,1-t)$ corresponds to $\max \sum_{j=1}^{i} I(n+1-p_j > i)$. The diagonal lines in the (U,V) plane, slopes ± 1, are used in exactly the same manner to move rectangles along and determine the rectangle with the largest volume under the density of (U,V). Rather than count points the population value seeks the rectangles with maximum volume under the density of (U,V).

For the specific bivariate normal for the figure, the correlation coefficient is $r = 3/(\sqrt{2}\sqrt{32}) = 3/8$. In formula (2) the supremums, as explained in the 1987 paper occur at t=1/2 and

$$C(\frac{1}{2},\frac{1}{2}) = \frac{1}{4} + \frac{1}{2p}\sin^{-1}\frac{3}{8} = 0.25 + 0.061179 = 0.311179, \text{ and}$$

$$C(\frac{1}{2},1) - C(\frac{1}{2},\frac{1}{2}) = \frac{1}{2} - (\frac{1}{4} + \frac{1}{2p}\sin^{-1}\frac{3}{8}) = 0.25 - 0.061179 = 0.18882.$$

Thus r_{gd} = 2(0.311179)-2(0.18882)=0.62235 – 0.37764= 0.2447.

For the bivariate Normal, the population value of r_{gd} is simplified to $\frac{2}{p}\sin^{-1} r$. So, directly, $\frac{2}{p}\sin^{-1}\frac{3}{8} = 0.2447$.

The volumes 0.3118 and 0.1888 are the volumes of the rectanglar regions along the diagonals emanating from the point (1/2,1/2) under the (U,V) density. In the figure of the (U,V) density the volume 0.3112 is the region under the raised part of the density , near the(0,0) point. The 0.1888 quantity came from the volume under that part of the density dropping off toward zero, near the point (1,0).

5. The Null Distribution of The GDCC

The null distribution for r_{gd} in the 1987 paper was given exactly up to sample size n = 10. Table 5 includes this original table and expands to list the probability distribution up to sample size n =15. The symmetry of the Null Distribution is used so that only probabilities of positive outcomes and zero are given. A critical value table was given in the original paper for hypothesis testing for alpha values of 0.10, 0.05, and 0.01. An interpolation term was given so that a randomized test could be used to obtain almost exact alpha levels. With the exact distributions for n =11 to 15, the simulation estimates can be replaced by more exact values. For example if n = 15 and a two-sided test of independence is desired, then from the original paper, to obtain a two-sided test with $a = 0.05$, reject H_o: independence, if $|r_{gd}| \geq 4/7$ and reject with probability 0.399 if $|r_{gd}| = 3/7$. The old probability was 0.36640. Thus,

$$P\left(reject\,|\mathrm{H}_{o}\ true\right)=P\left(\left|r_{gd}\right|\geq 4/7\right)+(0.399)P\left(\left|r_{gd}\right|=3/7\right)=$$
$$0.016474+(0.399)0.088309=0.05$$

Table 4: Null Distribution: Greatest Deviation Correlation Coefficient
frequency of outcomes, symmetric about zero
sample size

outcome	n=3	4	5	6	7	8	9	10	11
0	4	16	16	256	2848	11016	63720	1462104	14705496
1/5	0	0	0	0	0	0	0	562932	6664068
¼	0	0	0	0	0	11772	123660	0	0
1/3	0	0	0	196	500	0	0	0	0
2/5	0	0	0	0	0	0	0	479120	5128736
½	0	3	51	0	0	2480	18992	0	0
3/5	0	0	0	0	0	0	0	36672	732128
2/3	0	0	0	35	595	0	0	0	0
¾	0	0	0	0	0	399	6927	0	0
4/5	0	0	0	0	0	0	0	4623	80719
1	1	1	1	1	1	1	1	1	1
TOTAL	3!	4!	5!	6!	7!	8!	9!	10!	11!

sample size		
outcome	n = 12	13
0	83238912	1449824256
1/6	146029788	1509191388
1/3	32023332	633876372
½	19046768	216133376
2/3	727632	28456272
5/6	53823	940863
1	1	1
TOTAL	12!	13!

sample size		
outcome	n = 14	15
0	30683667456	330550419456
1/7	11736567360	269200784448
2/7	13614341004	150850248588
3/7	2117148516	57739685652
4/7	758459648	9645298832
5/7	20168080	1114989392
6/7	627263	10967359
1	1	1
TOTAL	14!	15!

This is just being saved for now:

Column 1 is headed by i, the won-lost ranking, Column 2 by p_i, the Sportmanship ranks, and column 4 by $n+1-p_i$, called the reverse ranks. The computations are shown in columns 3 and 5 and are the terms $\sum_{j=1}^{i} I(p_j > i)$, labeled I-sum-R, and the terms $\sum_{j=1}^{i} I(n+1-p_j > i)$, labeled I-sum-L, respectively. The hand, eye, brain algorithm for creating column 3 from columns 1 and 2 is exactly the same for column 5 using columns 1 and 4. Put fingers under the i of column 1 and under the corresponding positions on columns 3 and 5 and count the y ranks above that are strictly greater than i, and enter the result. Move down or increase i and repeat. With just a little practice, it is easy to perform this operation without really thinking of these formulae. the maximums used for GDCC are given at the bottom, 6 and 3 for this data. Thus $r_{gd} = -3/8$.