A Second Opinion Correlation Coefficient

Rudy A. Gideon and Carol A. Ulsafer

University of Montana, Missoula MT 59812

An important decision in most fields of research is whether two variables are related. Pearson's correlation coefficient usually answers this question, but it lacks robustness and depends on normality. Thus, inappropriate decisions can be made because of a few data points skewing the conclusion. This leads either to handling complex data by the dubious process of throwing out a selected set of the data or using other correlation coefficients. Spearman or Kendall can be used but neither seems to be as robust as the Greatest Deviation Correlation Coefficient, which gives a reliable "second opinion," as illustrated by the example presented.

1. INTRODUCTION

In the early 1980s Professor Rudy Gideon, the principal author of this paper, created a new correlation coefficient based on the ranks of the data and a counting method. He called it the Greatest Deviation Correlation Coefficient or GDCC. (For an example of the calculation of this correlation coefficient see Appendix A; for an introduction to the definition, some consequences and some examples, see Gideon and Hollister (1987).) He and his students have continued working with this new statistic ever since. While the body of knowledge grew in both depth and quantity, it has not been widely disseminated.

An entire theory of regression, including nonlinear and generalized linear models, starting with any correlation coefficient (rather than obtaining a correlation coefficient as a side calculation) as well as scale and location statistics have been developed. This approach has been extensively demonstrated with GDCC and somewhat with Kendall's Tau. (The interested reader is invited to sample these revolutionary ideas by reading the papers on the authors web site, www.math.umt.edu/gideon.)

This paper is purposefully expository and easy to read in order to focus on the idea that GDCC is extremely valuable in analyzing data.   We illustrate why more than the Pearson, Spearman, and Kendall CC's should be computed on bivariate data.

The data comes from the United States Department of Education and appeared in the International Edition of USA Today on December 20, 1984. While one may initially feel that this data set is too old to be relevant, the purpose of this paper is to illustrate a powerful statistical tool, not to draw inferences about a particular data set. However, data of this type appears regularly in today's world. This was the first large data set analyzed using GDCC; it was not contrived or sought after, but simply presented itself and most unexpectedly, showed the value of GDCC as a "second opinion," which has now become one of its primary uses. Subsequently, a large number of data sets have been analyzed and all results suggest that GDCC gives information above and beyond that of the classical three correlation coefficients.

2. ANALYSIS

The data is educational data from 50 states and Washington D.C. State averages on five variables related to high school education were recorded: (1) combined SAT scores, (2) high school graduation rate, (3) average teacher salary, (4) pupil-teacher ratio, and (5) average expenditures per pupil. In such data, researchers may be interested in seeing if there are relationships among these variables. This data cannot be considered independent and identically distributed because each state is distinctive; for example, no information on how the averages were made was given. Because of this, a researcher should want to treat each state with equal importance in searching for relationships; that is, no data should be discarded. The Greatest Deviation CC does treat data points on an equal basis and it will be seen that GDCC can find that some variables are related and can find that some are not in opposition to the decisions made by Pearson or Spearman or even Kendall. Thus, GDCC can help prevent both Type I and II errors. The data and the correlations appear at the end of this paper and should be read prior to the discussion. Table 1 shows the actual values of the correlation coefficients. In Table 2 are the transformation values which allow the NPCCs to be more directly compared to Pearson's correlation coefficient. This is a standard technique which can be done regardless of whether the data is actually bivariate normal, so that if the data is normal all correlation coefficients would estimate the same quantity. This is necessary because in general the NPCC are estimating correlation quantities smaller than the correlation parameter of the bivariate normal.

Experience has shown that good data is characterized by all these correlation coefficients having approximately the same significance. It must be emphasized that the correlation coefficients are estimating different quantities and only the significance levels should agree, not the values themselves. However, for problematic data, GDCC can have quite different significance levels than the others. Additionally, if the transformed GDCC as given in Table 2 has a value greater than the Pearson correlation coefficient, then this means that problematic data has devalued the Pearson correlation coefficient. This is so with or without significance. The situation is reversed in the opposite direction: if GDCC has a transformed value much less than Pearson's, a few points are inflating Pearson's. The example illustrates these concepts.

The Pearson, Spearman, Kendall, and GD correlation coefficients were computed on the ten pairs of variables and it was noted whether or not they were significant at the one and five percent levels. In the chart, the one percent significance level is denoted by double asterisks and the five percent by a single asterisk. Exact critical values were used for Pearson and GDCC and asymptotic values for Spearman and Kendall. For a data set with n=51, in order for GDCC to have an exact 5% critical point, one rejects the null hypothesis of independence (uncorrelated variables) if GDCC equals or exceeds 7/25 (Gideon and Hollister (1987)) and randomly rejects 58% of the time if it equals 6/25. For the 1% critical level, rejection is at 9/25 and rejection 76% of the time at 8/25.

In what follows the variables are referred to by their code numbers 1,2,3,4,5 defined above. All four computed correlation coefficients agreed completely on six of the relationships: variable pairs (1,4), (1,5), (2,3), (3,4) were not related by all measures (all four correlation coefficients showed nonsignificance) whereas (3,5) and (4,5) were related (all four correlation coefficients showed significance at the 1% level). For pair (2,4) there was partial agreement as all the NPCCs were more significant than Pearson. For pair (2,5) GDCC was more significant than the other three correlation coefficients, again giving partial agreement. However, a remarkable difference occurs for pairs (1,2) and (1,3). For (1,2), GDCC is not significant whereas the other three CCs are, but for pair (1,3), GDCC is significant at the 5% level while none of the others are. Which is right?

In trying to answer this question, let us compare the accepted procedure of deleting points and its effect on the analysis to using GDCC without the deletion of overly influential points. Even though it seems capricious to delete some states for some pairs of variables and different states for other pairs, this is probably the most common procedure in practice. To do this, one studies bivariate plots and uses influence measures to delete data of an unusual nature relative to the rest of the states. If this is done, the data for GA, MN, and WY are deleted for pair (1,3). This makes GDCC more significant and all the other CCs significant. Thus, three states masked a possible significant negative CC for three measures; GDCC was the exception. Reliance on GDCC would possibly avoid a Type II error without analyzing which data to remove. For pair (1,2) the deletion of DC, IA, MN, and SD makes GDCC even closer to zero and all three of the other CCs nonsignificant. In this case, three correlations are being made significant by just four states and only GDCC gave a result pointing in the correct direction both before and after deleting data points; using it possibly avoids a Type I error.   Note that different states were deleted for these two pairs, and hence, it is unclear what conclusion should be drawn for all the data with three of the CCs. In general, when psuedo outliers are deleted, most times the other correlational analyses now agree with the original GDCC results. GDCC makes the correlational analysis easy and allows reliable conclusions to be drawn. A valid analysis of a data set should be based on consistent use of the data.

Researchers have complex multivariate data and sometimes not a lot of time. While there are many other robust analyses, GDCC analysis is quick and easy and gives good second opinions; it also removes the necessity of deleting or weighting suspect data points. This example makes it clear how a small segment of the data can lead one to dubious conclusions. Since Least Squares estimation techniques are closely related to the Pearson's CC as shown in the simple regression paper available on the Web ("Correlation in Simple Linear Regression," www.math.umt.edu/gideon), it is clear that without a parallel robust NPCC analysis, many conclusions could be drawn some of which do not fairly represent the data.

Thus, in our example only GDCC pointed to a possible relationship between teacher salary and SAT scores, and it was negative. To amplify the differences the GDCC regression ("Correlation in Simple Linear Regression" on the web site) was run and gave

$$S\hat{A}T = 1130.84 - 0.008939 * teachersalary$$

whereas, Pearson's CC (slope and intercept same as least squares) gave

$$S\hat{A}T = 1008.27 - 0.002906 * teachersalary.$$

Note that for the 5% significant GDCC, its accompanying regression shows that an increase of $1000 in average teacher salary points to a decrease of 8.9 in average SAT score, but that Pearson's regression or least squares is nonsignificant and the corresponding decrease is only 2.9. The contradiction in higher salaries leading to lower SAT scores lends itself to interesting speculation; one conclusion might be that such data involving state averages shouldn't be used to draw inference about high school education.

3. BOOTSTRAP COMPARISON OF PEARSON AND THE GDCC

The two correlation coefficients, Pearson and the Greatest Deviation, were compared by running bootstrap analyses using SPLUS. These comparisons were done in order to connect this new information with the familiar and as expected they confirm the above comparsions. The GDCC, because it is nonparametric, is applicable under a wider set of assumptions than Pearson's correlation coefficient and so on that basis alone is more robust. However, the bootstrap confirms the two differences in inference about SAT and teacher salary and also about SAT and high school graduation rate, and in, addition, gives also the standard error and BCA (biased corrected and accelerated method).

First, the case of SAT and average teacher salary is considered. Recall that in this case GDCC is significant but the other three correlation coefficients are not. In this data the correlations were negative and the upper 5 % and 2.5% points reveal if there is significance. The upper bootstrap 5% and 2.5% points were -0.12 and -0.08 respectively for GDCC and 0.08 and 0.12 for Pearson; thus, Pearson's included zero in the confidence interval. The SEs on the mean were 0.1124 for GDCC which was less than the 0.1404 for Pearson.

The other case is that of SAT and high school graduation rate. Here the three well-known correlation coefficients are all significant but GDCC is not. The correlations are all positive so the lower bootstrap 2.5% and 5% points determine significance. The lower 2.5% and 5% points were -0.8 and -0.04 for GDCC, 0.14 and 0.18 for Pearson. The SEs for the mean were nearly the same: 0.1265 for GDCC and 0.1292 for Pearson.

4. CONCLUSION

Although only an example is presented here, numerous data sets have been examined over many years and the discerning character of GDCC as a valuable second opinion has held up. In a paper accepted for publication, "The Correlation Coefficients," the definitions of yet other correlation coefficients that could also provide better bivariate analysis are found. The asymptotic distribution and an area interpretation of GDCC appear in Gideon, Prentice and Pyke (1989); $\sqrt{n} * GDCC$ is asymptotically $N(0,1)$, but n should be at least 100 for a good approximation. GDCC is easy to compute by hand for small to medium sample sizes and examples appear in Appendix A and in Gideon and Hollister (1989). The SPLUS or R code for GDCC is given in Appendix B. It is given in two subroutines. The routine GDave is a function of the x and y data and returns GDCC. The routine rguniq is called by GDave. It gives the value of GDCC for a distinct set of y-ranks which correspond to the ordered x-data and is a function only on the y-ranks.

A unique — one method handles all cases — technique, rather than the usual local average rank method handles tied values. This method averages two calculated correlation coefficients and makes a value available for all cases and for many sorts of correlational analyses. For example, if all components of data vector x are the same, then this averaging method yields a GDCC value of zero, meaning no information rather than no relationship, because the two intermediate values are +1 and −1. These intermediates are the maximum and minimum possible correlation within the tied value restrictions. See Gideon and Hollister (1987). This is built into the computer programs in Appendix B. This tied value procedure can be used for all rank-based correlation coefficients, thereby making the calculations more consistent and not reliant on judgment calls.

5. APPENDIX A: A Hand Calculation of GDCC.

x ranks	y ranks	column 3	reverse y ranks	column 5
1	5	1	2	1
2	6	2	1	0
3	4	3	3	0
4	1	2	6	1
5	3	1	4	1
6	2	0	5	0
maxima		3		1

GDCC is calculated by example. The data are in columns 1 and 2, listed in order of the x ranks, and column 4 contains n+1− rank(y). For each x rank, a number in each of columns 3 and 5 is computed. At x rank 4 for example, 2 appears in column 3 because from {5,6,4,1}, the y ranks at or above the fourth x rank, only 5 and 6 (2 values) are strictly greater than the fourth x rank. There is a 1 in column 5 because from the reverse ranks, {2,1,3,6}, only one value (6) is strictly greater than 4. GDCC = (max(col 5) − max(col 3))/(greatest integer in n/2). Here this is (1−3)/3 = −2/3.

APPENDIX B: S-PLUS or R programs for the Calculation of GDCC

1. rguniq: Computes GDCC for the unique ranks of y where y has been sorted relative to x.

2. GDave: For any set of bivariate data, this routine computes two values of GDCC and averages them for a unique result. ccp is the value of GDCC computed so that ties are broken to achieve maximum positive correlation. ccn is the value of GDCC computed so that ties are broken to achieve the least positive correlation. Both ccp and ccn call rguniq.

```
1. rguniq <-
function(rky)
{ n <- length(rky); n1 <- n-1
   dy <- NULL; dyn <- NULL
   ryr <- n + 1 - rky
   for(i in 1:n1){
       dy <- c(dy, sum(rky[1:i] - i > 0))
       dyn <- c(dyn, sum(ryr[1:i] - i > 0))}
   mdyr <- max(dyn)
   mdy <- max(dy)

   corrg <- (mdyr - mdy)/(n %/% 2)
   corrg  }
```

```
2. GDave<-function(x,y)
{      n <- length(x)
    xt<-x[order(y,x)] #x order by y with y ties ordered by x
        rky<-1:n
        rky<-rky[order(xt,rky)]    # ranks of y ordered by x
        ccp  <- rguniq(rky)   # GD positive
# GD negative below

   xrr <-  n +1 -rank(x)  #reverse ranks on the x
   xt <- x[order(y,xrr)]  #x ordered by y with y ties ordered by rev(x)
   rky <- order(xt,n:1)   #ranks of y ordered by x with y ties
   ccn <- rguniq(rky)            #ordered by rev(y)

         (ccp+ccn)/2  }
```

6. REFERENCES

Gideon, R.A., and Hollister, R.A. (1987), "A Rank Correlation Coefficient Resistant to Outliers," *Journal of the American Statistical Association,* **82**, no.398, 656-666.

Gideon, Rudy A. (2007), "The Correlation Coefficients" (accepted for publication in *Journal of Modern Applied Statistical Methods*) and "Correlation in Simple Linear Regression" on the website www.math.umt.edu/gideon.

Gideon, R.A., Prentice, M.J., and Pyke, R.(1989), "The Limiting Distribution of the Rank Correlation Coefficient r_g", in *Contributions to Probability and Statistics* (Essays in Honor of Ingram Olkin), ed. Gleser, L.,J., Perlman, M.D., Press, S.J., and Sampson, A.R., New York: Springer-Verlang, pp. 217-226.

Kendall, M.G., and Gibbons, J.D. (1990), *Rank Correlation Methods*, 5th ed. Oxford University Press, or also Kendall, M.G. (1962), *Rank Correlation Methods*, 3rd ed. Hafner Publishing Company.

Pearson, K. (1911), "On The Probability That Two Independent Distributions Of Frequency Are Really Samples From The Same Population", *Biometrika,* **8**, 250-254.

Spearman, C. (1904), "The Proof And Measurement Of Association Between Two Things," *American Journal of Psychology*, **15**, 72-101

7. TABLES

Recall that the codes for the five variables are:
(1) combined SAT scores, (2) high school graduation rate, (3) average teacher salary, (4) pupil-teacher ratio, and (5) average expenditures per pupil.

Table 1:Correlations in the Education Data

variable	CC	2	3	4	5
1	Pearson	.4104**	-.1465	-.0878	-.1577
	Spearman	.3980**	-.1711	-.1672	-.1354
	Kendall	.2792**	-.1122	-.1208	-.0917
	GDCC	.1800	-.2400*	-.0400	-.0800
2	Pearson		.0970	-.2804*	.1614
	Spearman		.0824	-.4356**	.3056*
	Kendall		.0486	-.2752**	.2102*
	GDCC		.1400	-.3400**	.3200**
3	Pearson			-.0126	.8273**
	Spearman			.0891	.7336**
	Kendall			.0627	.5372**
	GDCC			.0400	.5600**
4	Pearson				-.4778**
	Spearman				-.4850**
	Kendall				-.3388**
	GDCC				-.3600**

NOTE: The Critical Values for the Correlation Coefficients in Table 1

Correlation Coefficient	Two-sided critical values (n= 51)	
	5% : one star (*)	1% : two stars (**)
Pearson	.279	.361
Spearman	.277	.364
Kendall	.188	.247
GDCC	7/25: 6/25 (.58)	9/25: 8/25 (.76)

Table 2: Transformations of the Correlations

variable	CC	2	3	4	5
1	Pearson	.4104	-.1465	-.0878	-.1577
	Spearman	.4137	-.1789	-.1748	-.1416
	Kendall	.4246	-.1753	-.1886	-.1435
	GDCC	.2789	-.3681	-.0628	-.1253
2	Pearson		.0970	-.2804	.1614
	Spearman		.0863	-.4522	.3186
	Kendall		.0763	-.4189	.3242
	GDCC		.2181	-.5090	.4817
3	Pearson			-.0126	.8273
	Spearman			.0932	.7494
	Kendall			.0983	.7472
	GDCC			.0627	.7705
4	Pearson				-.4778
	Spearman				-.5025
	Kendall				-.5074
	GDCC				-.5358

NOTE: If r is the population correlation coefficient for normally distributed data, then the following transformations make the correlation coefficients directly comparable:

Spearman: $\hat{r} = 2\sin(\frac{pr_s}{6})$ Kendall: $\hat{r} = \sin(\frac{pr_k}{2})$ GDCC: $\hat{r} = \sin(\frac{pr_{gd}}{2})$

Table 3: The Educational Data

State	SAT	graduation	teacher	pupil/teacher	expenditures
	(1)	rate (2)	salary (3)	ratio (4)	per pupil (5)
AL	970	64.4	17948	20.3	2177
AK	914	77.8	34510	13.2	7325
AZ	978	68.4	21119	19.5	2524
AR	1003	76.2	15310	18.2	1971
CA	897	75.1	23614	23.3	2733
CO	979	79.2	23276	18.6	3171
CT	904	77.9	21036	14.8	3636
DE	902	88.9	20625	17.5	3456
DC	823	58.4	25610	17.2	4260
FL	890	65.5	18275	17.8	2680
GA	822	65.9	13040	18.6	2169
HI	869	82.2	24319	22.9	3239
ID	992	77.9	17605	20.7	2052
IL	981	77.1	22972	18.0	3100
IN	864	78.3	20347	19.8	2414
IA	1089	88.0	19402	15.7	3095
KS	1051	82.5	18313	15.6	3058
KY	997	68.4	18384	20.2	2100
LA	980	57.2	18416	18.4	2739
ME	892	76.7	16248	19.5	2458
MD	897	81.4	22800	18.3	3445
MA	896	77.5	21841	16.1	3378
MI	976	73.4	25712	21.9	3307
MN	1020	90.7	22876	18.0	3085
MS	992	63.7	14320	18.6	1849
MO	981	76.2	17521	17.4	2468
MT	1034	83.1	19702	16.0	3289
NE	1041	84.1	17399	15.5	2984
NV	931	74.6	22067	20.9	2613
NH	931	76.5	16549	16.4	2750
NJ	876	82.7	21536	15.8	4007
NM	1014	71.4	20187	18.8	2901
NY	894	66.7	25000	18.8	4686
NC	827	69.3	17585	19.8	2162
ND	1054	94.8	18774	16.6	2853
OH	968	82.2	20004	19.8	2676
OK	1009	79.6	18270	17.0	2805
OR	907	73.0	21746	18.6	3504
PA	887	79.7	21178	17.2	3329
RI	885	75.2	23175	15.7	3570
SC	803	66.2	16523	18.9	2017
SD	1086	85.0	15592	15.5	2486
TN	1009	65.1	17698	20.9	2027
TX	886	69.4	19550	17.9	2731
UT	1045	84.5	19859	24.3	2013
VT	907	85.0	16299	13.9	3051
VA	894	75.7	18535	17.4	2620
WA	968	75.5	23485	21.7	3211
WV	976	77.4	17322	16.9	2764
WI	1007	84.0	21496	17.4	3237
WY	1034	81.7	23822	15.2	4045