Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 16

11-1-2007

The Correlation Coefficients

Rudy A. Gideon
University of Montana

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 517-529

1538 – 9472/07/$95.00

The Correlation Coefficients

Rudy A. Gideon
University of Montana

A generalized method of defining and interpreting correlation coefficients is given. Seven correlation coefficients are defined — three for continuous data and four on the ranks of the data. A quick calculation of the rank based correlation coefficients using a 0-1 graph-matrix is shown. Examples and comparisons are given.

Key words: Pearson, Spearman, Kendall, Gini, Greatest Deviation, median, absolute value, nonparametrics, correlation, tied values

Introduction

Definitions

This article introduces a system of estimation that has numerous advantages over current practice. Among these advantages is the global tied value procedure for nonparametric or rank based correlation coefficients making estimation functional over all data and advanced statistical methods, such as multiple regression; the currently used local tied value procedure is very restrictive. This system has produced a way of viewing correlation that has allowed other correlation coefficients to be defined. In particular, the new continuous absolute value and median correlation coefficients should be used for L1 methods or the MAD scale estimate. It is general and provides a robust estimation procedure in correlation analysis and in advanced statistical procedures if robust correlation is used (www.math.umt.edu/gideon).

Rudy Gideon received the Ph.D. in Statistics in 1970 under John Gurland at the University of Wisconsin. His academic career began in the Department of Mathematical Sciences at the University of Montana in 1970; he retired from the Department in June of 2005. He has worked extensively with Masters and Doctoral students as well as on a multitude of various applied statistical projects. His prime goal in retirement is to disseminate his original correlation estimation system that encompasses basic statistical methods.

To make the definitions of the correlation coefficients more natural, Pearson's r is reformulated. This re-expression of r also makes possible a natural definition of parametric and nonparametric correlation coefficients based on absolute values and medians. Let CC and NP stand for correlation coefficient and for nonparametric. Some NPCCs are defined based on counting techniques. A 0-1 graph-matrix is used to establish relationships. Finally, some data is analyzed to examine the relative robustness of the NPCCs .

Let $(x_i, y_i), i = 1,2,\ldots n$ be a bivariate data set. The usual mean notation will be used and $x_i^* = x_i - \bar{x}, y_i^* = y_i - \bar{y}, i = 1,2,\ldots n$ are the centered data. The sample covariance is proportional to $\sum x_i^* y_i^*$. To prepare for later definitions, this covariance is rewritten as

$$\sum x_i^* y_i^* = \left(\sum (x_i^* + y_i^*)^2 - \sum (x_i^* - y_i^*)^2\right)/4 .$$

In the uncentered notation, this can be written as

$$\left(\sum (x_i - \bar{x} + y_i - \bar{y})^2 - \sum (x_i - \bar{x} - y_i + \bar{y})^2 \right)/4 .$$

This form of the covariance function appeared as an interpretation of Pearson's r in Rodgers and Nicewater (1988), when their rescaled variance interpretations were added. Heuristic motivation

for this form as a measure of the relationship between the x-y data is now given and it holds for all CCs that are to be defined.

When there is positive correlation the terms $(x_i^* + y_i^*)^2 = (x_i - \bar{x} + y_i - \bar{y})^2$ will tend to be large, because the two deviations will tend to be in the same direction. The distance from a negative relationship is large, so the correlation would be positive. The terms $(x_i^* - y_i^*)^2, i = 1,2,\ldots n$ will have some canceling effect, so they will tend to be small. The net effect is that the covariance will be large. The distance from a positive relationship is small so that the correlation would be positive. When x and y are independent variables, a similar amount of canceling occurs in both terms and the covariance will fluctuate around zero. When there is negative correlation the distance from positive correlation will be large as the $(x_i^* - y_i^*)^2, i = 1,2,\ldots n$ terms will tend to be large, but cancellation will be occurring in the $(x_i^* + y_i^*)^2, i = 1,2,\ldots n$ terms, so the distance from negative correlation is small. Throughout this article the term distance does not mean just Euclidean distance, but is meant to describe the numerical measures of deviations from perfect positive or negative correlation.

These concepts are next elaborated in Euclidean n-space. For this paragraph *x* and *y* are the n-dimensional vectors of the centered data, normalized so that each has Euclidean length one, $\|x\| = \|y\| = 1$. Consider the vector x + *y* in n-space; the farther this vector is from the origin (for this vector the origin represents perfect negative correlation) the more positive is the correlation. For perfect positive correlation, cos(*x*, *y*) = 1 and $\|x + y\| = 2$; that is, distance from the origin is maximum. Consider the vector *x* – *y*. The closer this vector is to the origin, the more positive the correlation. For *x* – *y*, the origin represents perfect positive correlation and hence, $\|x - y\|$ small means distance from perfect positive correlation is small. Throughout this article the term distance does not mean just Euclidean distance, but is meant to describe the numerical measures of deviations from perfect positive or negative correlation.

To restate, for *x* – *y* the surface of the centered n-dimensional ball of radius 2 represents perfect negative correlation, so $\|x - y\|$ large means distance from perfect positive correlation is large. For perfect negative correlation, cos(*x*, *y*) = –1, and $\|x - y\| = 0$, so the distance from the ball of radius 2 is a maximum.

Another way to express this, in terms of parameters, is that there is positive correlation when V(X+Y) > V(X–Y) and negative correlation when the inequality goes in the other direction. The connection between distance away from negative correlation and V(X+Y) and also for distance away from positive correlation and V(X–Y) is now illustrated for a bivariate normal distribution.

Let Z_1 and Z_2 be standardized normal random variables with CC ρ. Note that $E(Z_1Z_2) = \rho = [V(Z_1+Z_2) - V(Z_1-Z_2)] / 4$. The term $V(Z_1+Z_2)$ equals distance from perfect negative correlation and is a linear function of ρ, namely $2 + 2\rho$. For $\rho = -1$ this distance is zero but for $\rho = +1$, this distance is 4. Similarly, $V(Z_1-Z_2)$ is distance from perfect positive correlation and it is $2 - 2\rho$. For $\rho = -1$, this distance is 4, but for $\rho = +1$, this distance is 0. Note that these distances are monotonic functions of ρ and the overall correlation $V(Z_1+Z_2) - V(Z_1-Z_2)$ combines to equal 4ρ. However, for some of the other correlation coefficients this combining of the distance measures does not simplify. Also note that in the case of Fisher's normal transformation,

$$\frac{1}{2}\ln\frac{V(Z_1+Z_2)}{V(Z_1-Z_2)} = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} =$$

$$\tanh^{-1}\rho = \ln\frac{\sqrt{V(Z_1+Z_2)}}{\sqrt{V(Z_1-Z_2)}}.$$

It is possible that a similar normalizing concept would work for other correlation coefficients.

Additionally, a correlation coefficient could be based on the ratio, V(X+Y) / V(X–Y), which would be less than one for negative correlation, one for independent random variables, and greater than one for positive correlation.

Pearson's r and other correlation coefficients based on absolute values and medians can now be defined. Let SS_X stand for a centered sum of squares and SA_X stand for the sum of absolute values about the mean; i.e.,

$$SA_x = \sum |x_i - \bar{x}| .$$

Continuous correlation coefficients
Definition 1: Pearson's r

$$r(x,y) = \frac{1}{4}\left(\sum \left(\frac{x_i^*}{\sqrt{SS_x}} + \frac{y_i^*}{\sqrt{SS_y}}\right)^2 - \sum \left(\frac{x_i^*}{\sqrt{SS_x}} - \frac{y_i^*}{\sqrt{SS_y}}\right)^2 \right) \quad (1)$$

= {(standardized distance from perfect negative correlation) – (standardized distance from perfect positive correlation)} divided by a constant, that puts the value between –1 and +1.

Definition 2: An absolute value CC, r_{av}

$$r_{av}(x,y) = \frac{1}{2}\left(\sum \left| \frac{x_i^*}{SA_x} + \frac{y_i^*}{SA_y} \right| - \sum \left| \frac{x_i^*}{SA_x} - \frac{y_i^*}{SA_y} \right| \right) \quad (2)$$

where y,i.e. $\sum \left| \frac{x_i^*}{SA_x} \right| + \sum \left| \frac{y_i^*}{SA_y} \right| = 2$

Definition 3: The Median Absolute Deviation correlation coefficient.

For the final continuous correlation, a correlation analog of the MAD, median absolute deviation estimate of variation, is given and denoted by r_{mad}. For a random sample, define $MAD_x = med|x_i - med(x_i)|$ and similarly for the data from Y. A median-type correlation coefficient is defined as

$$r_{mad} = \frac{1}{2}\left(med\left| \frac{x_i - med(x_i)}{MAD_x} + \frac{y_i - med(y_i)}{MAD_y} \right| - med\left| \frac{x_i - med(x_i)}{MAD_x} - \frac{y_i - med(y_i)}{MAD_y} \right| \right). \quad (3)$$

It is not true that $|r_{mad}| \le 1$. Let $x_i^* = \frac{x_i - med(x)}{MAD_x}$, and similarly for y_i^*. Now, $med|x_i^*| = med|y_i^*| = 1$.

The proof that $|r_{mad}| \le 1$ breaks down is because the median of the sum of two sets of nonnegative numbers is not always less than the sum of the medians. It would be true if the following equation held for r_{mad}.

$$med|x_i^* + y_i^*| \le med\left(|x_i^*| + |y_i^*|\right) \le med|x_i^*| + med|y_i^*| = 2$$

However, the second inequality does not always hold. The computer package S+ has been used to examine r_{mad}, and values slightly greater than one were occasionally obtained. Simulation studies of r_{mad} show it to behave very much like other correlation coefficients even with the anomaly of occasionally being greater than one. The spread of the distribution is very close to other correlations, and only when the population correlation is very near one can r_{mad} become slightly greater than one. In the case when X, Y have a bivariate normal distribution with parameters, $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho$, the population value is known to be $\rho_{mad} = \sqrt{\frac{1+\rho}{2}} - \sqrt{\frac{1-\rho}{2}}$. Substitute y for x in formula (3) and essentially MAD is recovered. Note that the same heuristic motivation for Pearson's r holds for this absolute value CC.

Rank based correlation coefficients

The first NPCC based on absolute values is now defined. In the same way that Spearman's CC is motivated from Pearson's r by using direct substitution of ranks, so is this new

correlation coefficient obtained from Definition 2 by substitution of ranks. An interesting historical note is that the NPCC in Definition 4 was found first and r_{av} determined from it.

First rewrite

$$(x_i - \bar{x}) + (y_i - \bar{y}) \text{ as } x_i + y_i - (\bar{x} + \bar{y}).$$

Replacing the data by their ranks and ordering the bivariate data by the x data, gives the data in rank form. this is just before the $(i, p_i), i = 1,2,\ldots n$. Thus p_i equals the rank of the y_i for the x with rank i. The means of the ranked data are $\frac{n+1}{2}$, so $\bar{x} + \bar{y}$ becomes $n+1$. The ranks p_i are here assumed distinct; tied values will be handled later. In Definition 2, with ranks substituted, the terms SA_x and SA_y are equal and can be factored from expression (2). Their value is

$$SA_x = SA_y = \sum \left| p_i - \frac{n+1}{2} \right| = \sum \left| i - \frac{n+1}{2} \right| = \sum \left| \frac{n+1-2i}{2} \right| .$$

For n odd, $\sum |n+1-2i|$ can be shown to be $\frac{n^2-1}{2}$ and for n even it becomes $\frac{n^2}{2}$; for either even or odd n, it is $\left[\frac{n^2}{2} \right]$, the greatest integer in $\frac{n^2}{2}$. Thus the denominator in (2) becomes

$$2SA_x = \sum |n+1-2i| = \left[\frac{n^2}{2} \right].$$

Definition 4: Spearman's modified footrule correlation coefficient, Gini (1914), Betro (1993)

$$r_{mf}(x,y) = \left(\sum |n+1-p_i-i| - \sum |p_i - i| \right) / \left[\frac{n^2}{2} \right] \quad (4)$$

The attempt by Spearman (1906) to make an absolute value rank CC was also documented in Kendall and Gibbons (1990). Spearman tried to make a computationally simple and robust CC and based it on one summation. The idea in this article is that all or at least most correlations should be a difference of two functions that measure distance from positive and negative correlation, which contrasts with Kendall's method in Chapter 2 in Kendall and Gibbons (1990). There, Kendall advanced the idea that some type of inner product should be used to define all CCs. The above two absolute value CCs cannot be defined using Kendall's inner product concept. This difference of two functions gives the necessary symmetry to a CC. The denominator arises from the absolute value of the numerator which occurs when $p_i = i$ (correlation = +1), or when $p_i = n+1-i$ (correlation = –1). Note again that the same heuristic motivation applies. The forumulation of Spearman's correlation coefficient based on Definition 1 is:

Definition 5: Spearman's correlation coefficient, Spearman (1906)

$$r_s(x,y) = \frac{n(n^2-1)}{3} \left(\sum (n+1-p_i-i)^2 - \sum (p_i - i)^2 \right) \quad (5)$$

$$= 1 - \frac{6}{n(n^2-1)} \sum (p_i - i)^2 .$$

The linear restriction that allows r_S to simplify as shown does not hold for r_{mf}. Two more CCs are to be defined — Kendall's, for which a linear restriction does allow a simplification of the defining formula and one based on maximum or greatest deviations for which no simplification occurs. Again the natural definitions are based on the difference of

two functions that measure distance from perfect positive and negative correlation and makes the distribution of the CCs symmetric about zero for the case when x and y are independent, i.e. the null case. It will also be shown that r_{mf} can be computed from the quantities defined for the numerator of the Greatest Deviation CC.

Both Kendall's CC (r_k), usually called Tau, and the one based on greatest deviations (r_{gd}) use a counting technique that can be defined with an indicator function. Let

$$I(.) = \begin{cases} 1 \text{ if the argument is true} \\ 0 \text{ if false} \end{cases}$$

Recall that the data are assumed ordered by the x data and for the i^{th} largest element of x, the rank of the corresponding y data is p_i. For Kendall's correlation coefficient, let

$$\sum_{j=i+1}^{n} I(p_j > p_i) = n_{c,i}$$

count the number of concordances and

$$\sum_{j=i+1}^{n} I(p_j < p_i) = n_{d,i}$$

count the number of discordances at position i (recall that no tied values are yet allowed). The larger the number of concordances the smaller the number of discordances. Let n_c and n_d be the sum over i, i=1,2,..., n-1 of the concordances and discordances, respectively. The concordance function, n_c, is a counting function that measures distance of the ranked data from a perfect negative monotone relationship, whereas n_d is a similar discrete measure of the ranked data from a perfect positive monotone relationship.

Definition 6: Kendall's r_k correlation coefficient, see e.g. Kendall and Gibbons (1990)

$$r_k(x,y) = \left(\sum_{i=1}^{n-1} n_{c,i} - \sum_{i=1}^{n-1} n_{d,i} \right) / \binom{n}{2} \qquad (6)$$

$$= (n_c - n_d) / \binom{n}{2}$$

$$= (4n_c / (n(n-1))) - 1 =$$

$$1 - (4n_d / (n(n-1))),$$

because $n_c + n_d = \binom{n}{2}$.

The quantity $\binom{n}{2}$ means n choose 2. This summation of n_c and n_d will be shown in the next section to be n choose 2 using a 0-1 graph-matrix formulation of the calculation of r_k.

For the Greatest Deviation CC let $d_i^+ = \sum_{j=1}^{i} I(p_j > i)$, a function that is large when there is negative correlation and small if not; that is, the measure is large if distance from positive correlation is great. Let

$$d_i^- = \sum_{j=1}^{i} I(n+1-p_j > i).$$

This is a measure that is large if distance from negative correlation is great.

Definition 7: The Greatest Deviation correlation coefficient, r_{gd}; Gideon and Hollister (1987) and in Gideon, Prentice, and Pyke (1989)

$$r_{gd}(x,y) = (\max_{1 \le i \le n} d_i^- - \max_{1 \le i \le n} d_i^+) / \left[\frac{n}{2} \right] \qquad (7)$$

where $\left[\frac{n}{2} \right]$ is the greatest integer in n/2; it's value is the maximum value of the difference in the numerator.

This completes the definitions of the correlation coefficients under consideration. The next section gives some insightful examples; the work is considerably eased using a computational aid that allows computations of the four nonparametric correlation coefficients

from an augmented plot of the data with a 0-1 matrix, called a graph-matrix.

Methodology

Computations using the graph-matrix

The data in rank form are $(i, p_i), i = 1,2,\ldots n$. Let e = (1,2,....,n) and p = $(p_1, p_2, \ldots, p_n)$ be the data in vector form. The graph of the ranked data will have e plotted on the horizontal axis and p plotted on the vertical axis.

The YMCA basketball data that were used in illustrating the Greatest Deviation CC (Gideon & Hollister, 1987) is used here again. These data occurred as ranks and they will now be used to calculate all four of the NPCCs that have been defined. The e contains the ranks of the won-lost records of the 16 teams that were in the fifth grade league in Missoula, Montana in 1980. Rank one is the team with the best record. Throughout the season, after each game, each coach was asked to rate the sportsmanship of the opposing team and at the end of the season the cumulative ratings were presented in rank form with rank one being the team with the highest rated sportsmanship. These ranks were p = $(14,11,16,2,12,13,7,9,10,3,8,1,15,6,4,5)$.

Note that in general the teams with the best won-lost records had the lower sportsmanship ratings. The correlation coefficients put a measure on the relationship between winning and sportsmanship.

The graph-matrix appears in the middle of Figure 1 surrounded by auxiliary information. The two leftmost and the two rightmost columns as well as the two bottom rows are intermediate calculations explained below. Bordering the data plot are the axes labels. The *s indicate the plotted points $(i, p_i), i = 1,2,\ldots n$ and unlike a scatterplot, the Cartesian product, e x e, on the graph is filled in with 0s above each of the plotted points and 1s below. The combination of these *s, 0s, and 1s are used to calculate all four NPCCs which appear on the three borders.

Although the definitions of the correlation coefficients may seem unwieldy, the counting technique is easy and quick to use. It is really more convenient to use the method if the diagonals to the data plot are drawn in, which is easier done by hand. The line of slope one is denoted sl^{+1}; this is the line through $(i,i), i = 1,2,\ldots,n$. The line of slope minus one, sl^{-1}, goes through points $(i, n+1-i), i = 1,2,\ldots,n$.

Immediately below the graph are two rows that give the values necessary to calculate the Spearman and Absolute Value CCs. The upper row counts from the * to the line sl^{-1} with a minus sign if the * is below sl^{-1}. The lower row counts from the * to the line sl^{+1} again with a minus sign if the * is below the line. It is readily apparent that this counting technique directly corresponds to the summands in the formulas of Definitions 4 and 5. The sum of the absolute values of these two rows are given just to the right of them (56, 106), followed by the sum of squares of them (348, 1012).

To the right of the graph-matrix are two columns that give the individual concordances and discordances in Kendall's Tau as given in Definition 6. Starting at a * in position (i, p_i), a 0 appears in column j>i (to the right of the *) if and only if the rank of that column p_j is in discordance ($p_i > p_j$) and a 1 appears in a column to the right of the * if and only if the rank of that column is in concordance ($p_i < p_j$). To obtain the discordances, count the 0s to the right of the * in each column, and to obtain the concordances count the 1s to the right of each * in each column. These results appear in the two columns to the right of the graph. The sums of the two columns, the total numbers of con- and discordances, are given below the columns as (38, 82). Note that the ordering within the two columns does not match the standard algorithm used to calculate Kendall's Tau, r_k.

To the left of the graph are two columns headed by d_i^+ and d_i^-. They label the values for which the maximums need to be taken in Definition 7 of the Greatest Deviation correlation coefficient. For each element in the d_i^- column count all the 0's on and to the left of

YMCA basketball data: correlation computations

left: Greatest Deviation bottom: Spearman and Absolute Value right: Kendall

d_i^+	d_i^-	vertical axis: sportsmanship rankings horizontal axis: won and lost standings																	n_c	n_d
0	1	16	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
1	2	15	0	0	1	0	0	0	0	0	0	0	0	0	*	0	0	0	0	3
2	1	14	*	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	2	13
3	2	13	1	0	1	0	0	*	0	0	0	0	0	0	1	0	0	0	1	9
3	2	12	1	0	1	0	*	1	0	0	0	0	0	0	1	0	0	0	2	9
4	1	11	1	*	1	0	1	1	0	0	0	0	0	0	1	0	0	0	4	10
5	2	10	1	1	1	0	1	1	0	0	*	0	0	0	1	0	0	0	1	6
6	2	9	1	1	1	0	1	1	0	*	1	0	0	0	1	0	0	0	2	6
6	2	8	1	1	1	0	1	1	0	1	1	0	*	0	1	0	0	0	1	4
5	2	7	1	1	1	0	1	1	*	1	1	0	1	0	1	0	0	0	4	5
5	2	6	1	1	1	0	1	1	1	1	1	0	1	0	1	*	0	0	0	2
4	3	5	1	1	1	0	1	1	1	1	1	0	1	0	1	1	0	*	0	0
3	3	4	1	1	1	0	1	1	1	1	1	0	1	0	1	1	*	1	1	0
3	2	3	1	1	1	0	1	1	1	1	1	*	1	0	1	1	1	1	5	1
2	1	2	1	1	1	*	1	1	1	1	1	1	1	0	1	1	1	1	11	1
1	0	1	1	1	1	1	1	1	1	1	1	1	1	*	1	1	1	1	4	0
6	3	gd	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	38	82
53	28	mf	-2	-4	2	-11	0	2	-3	0	2	-4	2	-4	11	3	2	4	56	348
			13	9	13	-2	7	7	0	1	1	-7	-3	-11	2	-8	-11	-11	106	1012

Figure 1.

the sl^{-1} line. To obtain each element in the d_i^+ column count all the 1's on and to the left of the sl^{+1} line. For example, $d_7^+ = \sum_{j=1}^{7} I(p_j > 7) = 5$, because exactly p_1, p_2, p_3, p_5, p_6 are greater than 7.

Using the graph, there are exactly 5 1s on or to the left of sl^{+1} in row 7, corresponding precisely to the five $p_i's$ mentioned above, because in that part of the plane, the second coordinate exceeds the first. Similarly, $d_7^- = \sum_{j=1}^{7} I(p_j < 17 - 7 = 10) = 2$ because only p_4 and p_7 are less than 10. Now for d_i^-, the term $n+1-p_j > i$ in the indicator function means $p_j < n+1-i$; that is, count all the zeroes at $n+1-i$ on the vertical axis on and to the left of the sl^{-1} line. So for i=7, count all the zeroes at 17-7=10 on the vertical axis on and to the left of sl^{-1}; the 0s appear only in columns 4 and 7 corresponding to p_4 and p_7 being less than 10.

Just below the d_i^- and d_i^+ columns are the maximums for r_{gd} and the below them are the sums of these two columns. It will be shown that these sums can be used to compute r_{mf}.

Note that twice 53 is 106 and twice 28 is 56, the numbers needed for r_{mf}.

From the statistics given in Figure 1, the differences in the numerators of the four correlation coefficients can be obtained and the denominators are

$$\left[\frac{n^2}{2}\right] = 128,\ n(n^2-1)/3 = 1360,\ \binom{n}{2} = 120,$$

$$\left[\frac{n}{2}\right] = 8.$$

$$r_{mf} = \frac{56-106}{128} = \frac{-25}{64} = -0.3906,$$

$$r_s = \frac{348-1012}{1360} = \frac{-83}{170} = -0.4882$$

$$r_k = \frac{38-82}{120} = \frac{-11}{30} = -0.3667,$$

$$r_{gd} = \frac{3-6}{8} = \frac{-3}{8} = -0.3750.$$

Note that the two numbers in the numerator for r_S and r_K add to the denominator (r_S: 348 + 1012 = 1360, and r_K: 38 + 82 = 120), the well-known linear restriction, but this does not occur for r_{gd} and r_{mf} as r_{gd}: 6 + 3 = 9 > 8, and r_{mf}: 56 + 106 = 162 > 128.

Special form for calculation of r_{gd}

If only r_{gd} is desired, there is a convenient algorithm to compute the d_i^- and d_i^+ values. Write down for i = 1,2, ..., n the three rows vectors $(i, p_i, n+1-p_i)$. Compute d_i^+ by placing a marker just to the right of the i^{th} position and count left in the p_i row and note all the ranks greater than i. Compute d_i^- by keeping the same marker, but counting left in the $n+1-p_i$ row noting all the ranks greater than i. This is done in Table 2. Note that d_i^- in Figure 1 and Table 2 appear in the same order whereas, the d_i^+ values are reversed.

Three theorems are given below which show some additional usefulness of this graph-matrix approach. The first shows the relationship between the statistics used in r_{gd} and r_{mf}.

Theorem 1: $2\sum d_i^+ = \sum |p_i - i|$ and $2\sum d_i^- = \sum |n+1-p_i-i|$, all sums from 1 to n.

Proof: First the d_i^+ relationship is established. Clearly $\sum_{i=1}^{n}(p_i - i) = 0$; that is, the sum of the deviations about the sl^{+1} is zero. Thus, $-\sum_{p_i<i}(p_i - i) = \sum_{p_i>i}(p_i - i)$. Now $\sum_{p_i>i}(p_i - i)$ just counts all the 1s on or above the sl^{+1} line. But, $d_i^+ = \sum_{j=1}^{i} I(p_j > i)$ counts all the 1s in row I that are on or above the sl^{+1} line so that $\sum d_i^+ = \sum_{p_i>i}(p_i - i) = \sum_{p_i<i}(i - p_i)$ or

$$2\sum d_i^+ = 2\sum_{p_i>i}(p_i - i) = \sum_{i=1}^{n}|p_i - i|.$$

These equalities are demonstrated in Figure 1. The bottom two rows carry signs to allow these equalities to be easily seen. The proof of the d_i^- relationship follows in a similar manner.

Theorem 2: The number of 1s on or to the right of the sl^{-1} line in row i-1 equals the number of 0s on or to the left of sl^{-1} in row i, i=2,3,...,n. The number of 0s on or to the right of the sl^{+1} line in row i equals the number of 1s on or to the left of the sl^{+1} line in row i-1, i=2, 3,...,n. (In this theorem row i refers to the vertical axis, which are ranks; e.g. row 1 corresponds to the bottom row of the 0-1graph-matrix.) Figure 1 provides a guideline for the proof.

The symmetry displayed in this theorem shows that the Greatest Deviation CC could have been equivalently defined in a right-handed fashion; i.e. instead of counting 0s and 1s from the left to the diagonal lines, counting could have been done from the right with a suitable adjustment.

Theorem 3: For Kendall's CC, $n_c + n_d = \binom{n}{2}$.

Proof: If the data positions (*s) fell on the diagonal of the graph-matrix it is clear that there would be a total of $n^2 - n$ 0s and 1s with complete anti-symmetry. The permutation of the columns to depict the actual data does not change this total and hence, the total number of 0s and 1s to the left of the *s must equal the total number to the right. Thus, $n_c + n_d = \frac{n^2 - n}{2} = \binom{n}{2}$. Further, the number of 1s to the right (38 in Figure 1) equals the number of 0s to the left and the number of 0s to the right (82 in Figure 1) equals number of 1s to the left.

Results

Which correlation coefficients are outlier resistant? In this section two examples are given to illustrate that the four NPCCs can have quite different values on the same data. The maximum differences between r_k and r_s appear on page 34 of Kendall and Gibbons (1990). The examples below suggest that r_{gd} and r_{mf} are the most robust, r_k next, but that Spearman's r_s is not very robust. Let e and p be the rank vectors. The calculation of the correlation coefficients is left to the reader. The values of the NPCCs for n = 10 and $p = (5,4,3,2,1,10,9,8,7,6)$ are

$$r_{mf} = \frac{26}{50} = 0.5200,\ r_s = \frac{17}{33} = 0.5152,$$
$$r_k = \frac{1}{9} = 0.1111,\ r_{gd} = \frac{3}{5} = 0.6000.$$

The values of the CCs now with $p = (10,1,2,3,4,5,6,7,8,9)$ are

$$r_{mf} = \frac{14}{50} = 0.2800,\ r_s = \frac{6}{330} = 0.0182,$$
$$r_k = \frac{11}{45} = 0.2444,\ r_{gd} = \frac{3}{5} = 0.6000.$$

It is known that for the bivariate normal distribution, the NPCCs estimate a function that is less than the correlation parameter, ρ. When the CCs differ greatly, it suggests that there are strange observations in the data. Here, r_{gd} and r_{mf} give the largest indication of a positive

Table 2. Calculation of the Greatest Deviation CC

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	max
p_i	14	11	16	2	12	13	7	9	10	3	8	1	15	6	4	5	
$n+1-p_i$	3	6	1	15	5	4	10	8	7	14	9	16	2	11	13	12	
d^+_i	1	2	3	3	4	5	5	6	6	5	4	3	3	2	1	0	6
d^-_i	1	2	1	2	2	1	2	2	2	2	2	3	3	2	1	0	3

Table 3 Spearman data

Person	addition	sound
D	1	3
I	2	2
H	3	1
B	4	4.5
J	5	4.5
E	6	11
A	7	6
K	8	9
F	9	8
C	10	10
G	11	7
	(i)	(p_i)

relationship for the strange data of these two examples. Hence, they may be the most resistant to outliers or to any unusual data. (Work in progress shows them more resilient.)

Probabilities and asymptotics for the rank correlation coefficients

Some aspects of the rank CCs will be compared by using an example from Spearman (1906) concerning the relationship between the ability of people to add numbers quickly and accurately and their ability to distinguish between two sound tones. Spearman used this example to illustrate his footrule CC. The data were for eleven students of psychology; Spearman ranked their ability in pitch discrimination and a second person ranked independently for addition ability. The data are ordered by the addition variable and note the two tied values with the usual convention used, which could be called a local convention as opposed to a more useful global definition given below.

Spearmans's footrule CC is

$$r_f = 1 - \frac{6\sum_{p_i > i}(p_i - i)}{n^2 - 1} = 1 - \frac{6(8.5)}{120} = 0.57 .$$

Because this footrule only involved distance from perfect positive correlation, it is not a valid correlation coefficient. It is interesting from a historical perspective. He compared this number to probable error (derived in his article) of 0.13 and concluded because 0.57/0.13 = 4.38, "the faculty of adding numbers and that of discriminating pitch is just about large enough to be beyond all reasonable suspicion of mere chance coincidence" (p. 96).

Spearman did not use a table of critical values but instead stated a heuristic value for the above ratio to be significant. The four nonparametric CCs and their corresponding probability values are now computed for this data. Referring to what is now known as the Spearman CC (the rank equivalent of Pearson's CC; i.e., r_S) Spearman said, "the effect of squaring is to give more weight to the extreme differences as compared with the median ones. This is probably a considerable advantage in most physical measurements. But in other fields of research, and perhaps above all in Psychology, these extreme cases are just the ones of most suspicious validity, so that the squaring is here more likely to do harm than good" (p. 99). Thus, Spearman wanted a robust CC for his data.

This example illustrates the definition of a rank CC when tied values are present. In

advanced work on the use of CCs in estimation, the current local methods of tied value calculations are not adequate and hence a global method first introduced in Hollister and Gideon (1987) is presented. In this method, the calculations are done twice: first when Person B is assigned rank 4 for sound and Person J is assigned rank 5 for sound, favoring positive correlation; in the second calculation ties are broken in the reverse direction to favor negative correlation. Note that r_{gd} is the only CC without a change. Each CC can be defined uniquely by averaging the values of the two extreme correlation coefficients.

In Table 4, r_{gd} remains at 0.6000 but r_{mf} becomes (0.7333 + 0.7000)/2 = 0.7167. A general global definition for an alternative tied value procedure is now given.

Definition: The global values of rank CC when ties are present

Let (x, y) be a set of data, and (I, P^+) be the corresponding ranks which are assigned among the tied values in the way that most favor positive correlation, and let (I, P^-) the corresponding ranks assigned among the tied values in the way to most favors negative correlation. I becomes e and P^+ and P^- are permutations of e. Then a rank correlation coefficient, r, is defined uniquely from the two extremes, P^+ and P^-. Its value is

$$r(x, y) = (r(e, P^+) + r(e, P^-))/2. \quad (8)$$

The quantities $r(e, P^+)$ and $r(e, P^-)$ are abbreviated to r^+ and r^-, respectively. As an example, let (x, y) = ((1,2,2,4,5), (1,1,2,1,3)). Then P^+ = (1,2,4,3,5) and P^- = (3,4,2,1,5). Thus, for r_{gd},

$$r_{gd} = \frac{r^+ + r^-}{2} = \frac{1/2 + (-1/2)}{2} = 0.$$

Return to the level of significance for the Spearman example. The numerators and values of the four NPCCs as computed by the 0-1 graph-matrix method are given in Table 4. The denominators are

$$\left[\frac{11}{2}\right] = 5, \binom{11}{2} = 55, \left[\frac{11^2}{2}\right] = 60,$$
$$\frac{11(11^2 - 1)}{3} = 440.$$

Tail probabilities are obtained from Neave (1978) for r_k and r_s, from Gideon and Hollister (1987) for r_{gd}, and from Betro (1993) for r_{mf}. The table values are compared to the asymptotic values computed from the asymptotic distributions which are given in Kendall and Gibbons (1990) for r_s and r_k and in Gideon, Prentice, and Pyke (1989) for r_{gd}. The asymptotic null distributions ($\rho = 0$) of the four CCs are given first. These are

$$\sqrt{n-1}r_s \text{ is } N(0,1); \sqrt{n-1}r_k \text{ is } N(0,4/9);$$
$$\sqrt{n}r_{gd} \text{ is } N(0,1); \sqrt{n-1}r_{mf} \text{ is } N(0,2/3).$$

For completeness the exact variances of each CC is given; $V(r_s) = 1/(n-1)$; $V(r_k) = 2(2n+5)/(9n(n-1))$; $V(r_{gd})$ is unknown; $V(r_{mf}) = 2(n^2+2)/(3n^2(n-1))$ for n even and $2(n^2+3)/(3(n-1)(n^2-1))$ for n odd. The one tie is neglected and the data for the most correlation case, P^+, is used. First, from tables,

$$0.001 \le P(r_s \ge 0.7636) \le 0.005;$$
$$0.01 \le P(r_k \ge 0.5636) \le 0.025;$$
$$0.01 \le P(r_{gd} \ge 0.6000) \le 0.05;$$
$$P(r_{mf} \ge 11/15 = 0.7333) = 0.0013 \text{ and}$$
$$P(r_{mf} \ge 7/10 = 0.7000) = 0.0024.$$

Thus, all of the CCs are significant with r_s and r_{mf} being the most significant. These results are now compared to the asymptotic approximations using the notation of Z as N(0,1).

Table 4.

Spearman's 1906 Data and Correlations
The pairs of numbers in the numerators show distances from – and + correlation
Correlations are in second row of the named correlation

	most +	most -	average
r_{gd}	5-2	5-2	
	0.6000	0.6000	0.6000
r_{mf}	60-16	60-18	
	0.7333	0.7000	0.7167
r_k	43-12	42-13	
	0.5636	0.5273	0.5455
r_s	388-52	386-54	
	0.7636	0.7545	0.7591

Table 5: Some asymptotic comparisons

$$P(r_s \geq 0.7636) \cong P(Z \geq \sqrt{10}(0.7636) = 2.4147) = 0.0079$$

$$P(r_k \geq 0.5636) \cong P(Z \geq \frac{\sqrt{10}(0.5636)}{2/3} = 2.6734) = 0.0038$$

$$P(r_{gd} \geq 0.6000) \cong P(Z \geq \sqrt{11}(0.6000) = 1.9900) = 0.0233$$

$$P(r_{mf} \geq 0.7333) \cong P(Z \geq \frac{\sqrt{10}(0.7333)}{\sqrt{2/3}} = 2.8401) = 0.0023$$

All of these approximate results are reasonably good. All four correlations support Spearman's conclusion that his footrule CC gave. Spearman drew his conclusion by comparing his footrule value of 0.57 to the probable error, which he gave as 0.13. Thus, 0.57/0.13 = 4.38. This example is concluded by comparing the value of r_{mf}, the modified footrule CC, 0.7333, to

$$\sqrt{V(r_{mf})} = \sqrt{\frac{2(11^2+3)}{3(10)(11^2-1)}} = \sqrt{0.0689} = 0.2625 .$$

Now, 0.7333/0.2625 = 2.7937 and by Spearman's rule of "satisfactory demonstration" that this ratio be at least 4, had Spearman found the correct formulation, r_{mf} , he would have drawn the opposite conclusion (p. 96).

Again for this example it should be pointed out that r_S and r_k have a linear restriction but r_{mf} and r_{gd} do not. Hence, the terms in the numerator, when added give the denominator for r_S and r_k but not for r_{mf} and r_{gd}. For r_S: 388+52 = 440 and for r_k: 43+12 = 55 whereas for r_{mf}: 60+16 = 66 > 60 and for r_{gd}: 5+2 = 7 > 5.

Conclusion

By viewing correlation broadly as the difference between measures of distance from perfect negative and perfect positive correlation, many new formulations of correlation may be defined. Two new continuous correlation coefficients are based on absolute values and medians. The median one is an extension of the MAD scale measurement and the absolute value one produces Gini's CC when data ranks are substituted. A 0-1 graph-matrix was introduced as an extension to the plot of the bivariate rank data and used to compute all four nonparametric correlation coefficients and exhibit some relationships. Several examples suggest which of the correlations are most robust: the Greatest Deviation and Gini. A data set from Spearman was used to demonstrate the application of the asymptotic distributions, to compare the correlations on the same data, and to illustrate a global tied value procedure. This procedure does not seem critical here, but for later developments on the use of correlation coefficients in estimation it is essential. Several times the normal distribution was selected to set up notation but this is not necessary, as any distribution from the class of bivariate t distributions would suffice. The four nonparametric correlation coefficients would be distribution-free on this class of bivariate distributions with elliptical shaped contours, including the Cauchy distribution.

.

References

Betro, B. (1993), On the Distribution of Gini's Rank Correlation Association Coefficient, *Communications in Statistics: Simulation and Computation*, *22*, No. 2, 497-505.

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

Gideon, R. A., Prentice, M. J., & Pyke, R. (1989). The Limiting Distribution of the Rank Correlation Coefficient r_{gd}, appears in *Contributions to probability and statistics* (Essays in Honor of Ingram Olkin) edited by Gleser, L. J., Perlman, M. D., Press, S. J., & Sampson, A. R. NY: Springer-Verlang, p 217-226.

Gideon, R. A. www.math.umt.edu/gideon.

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

Kendall, M. G. & Gibbons, J. D. (1990), *Rank correlation methods*, 5th ed. Oxford University Press, or also Kendall, M. G. (1962), *Rank correlation methods*, 3rd ed. GB: Hafner Publ. Co.

Neave, H. R. (1978), *Statistical tables*, London: George Allen & Unwin Publishers, Ltd/.

Rodgers, J. L. & Nicewater, W. A. (1988), Thirteen Ways to Look at the Correlation Coefficient, *The American Statistician*, *42*, no. 1, 59-66.

Rousseeuw, P. J. & Croux, C. (1993), Alternatives to the Median Absolute Deviation, *Journal of the American Statistical Association*, *88*, 1273-1283.

Scarsini, M. (1984), On Measures of Condordance, *Stochastica*, *8*, No. 3, 201-218.

Schweizer, B. & Wolfe, E. F. (1981), On Nonparametric Measures of Dependence for Random Variables, *The Annals of Statistics*, *9*, 879-885.

Spearman, C. (1906), 'Footrule' for Measuring Correlations, *British Journal of Psychology*, *2*, 89-108.