Chapter 1: **THE CORRELATION COEFFICIENTS**

INTRODUCTION

In this chapter there are seven correlation coefficients defined; three for continuous data and four on the ranks of the data. Two of the continuous correlation coefficients are new and based on absolute values and medians and no extensive study has yet been made on them. The median one appears as an exercise. The other continuous one is the well-known Pearson correlation coefficient but it is expressed in a new form to allow a new interpretation. Three of the rank correlation coefficients have appeared in print but only the Spearman and Kendall ones are well-known; the unknown one is based on the principle of maximum or greatest deviation and has been well studied. Because it is not a linear function, it has forced a more general way to look at the many possible uses of correlation coefficients. The last rank correlation coefficient is defined with absolute values and is called the Spearman Footrule correlation coefficient; it has been studied somewhat and is the nonparametric counterpart of the new continuous correlation coefficient based on absolute values.

Section 1: BASIC DEFINITIONS

Before defining the correlation coefficients, Pearson's r will be expressed in a form that will make the definitions more natural. This re-expression of r will also make possible a natural definition of parametric and nonparametric correlation coefficients based on absolute values. For now let CC and NP be abbreviations for correlation coefficient and nonparametric. Some NPCC's will be defined based on counting techniques and a 0-1 matrix will be used to easily establish certain relationships. Finally, some data will be analyzed to examine the relative robustness of the NPCC's .

Let $(x_i, y_i), i = 1, 2, \cdots, n$ be a bivariate data set. The usual mean notation will be used and $x_i^* = x_i - \bar{x}, y_i^* = y_i - \bar{y}, i = 1, 2, \cdots, n$ 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 $\sum x_i^* y_i^* = (\sum (x_i^* + y_i^*)^2 - \sum (x_i^* - y_i^*)^2)/4$. In the uncentered notation, this can be written as $(\sum (x_i - \bar{x} + y_i - \bar{y})^2 - \sum (x_i - \bar{x} - y_i + \bar{y})^2)/4$. This form of the covariance function appears as an interpretation of Pearson's r in Rodgers and Nicewater (1988), when their rescaled variance interpretations are added together. Some heuristic motivation for this form as a measure of the relationship between the x-y data is now given and it will hold for all CC's that are to be defined. When there is a positive correlation the terms $(x_i^* + y_i^*)^2 = (x_i - \bar{x} + y_i - \bar{y})^2, i = 1, 2, \cdots, n$ will tend to be large as the two deviations will tend to be in the same direction. One could say that the "distance" from a negative relationship is large so that the correlation would be positive. On the other hand, the terms $(x_i^* - y_i^*)^2, i = 1, 2, \cdots, n$ will have some cancelling effect so that they will tend to be small and the net effect is that the covariance will be large. One could say that 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 cancelling occurs in both terms and the covariance will fluctuate around zero. When there is a negative correlation the "distance"

from positive correlation will be large as the $(x_i^* - y_i^*)^2, i = 1,2,\cdots,n$ terms will tend to be large but cancellation will be occuring in the $(x_i^* + y_i^*)^2, i = 1,2,\cdots,n$ terms, so that the "distance" from negative correlation is small.

We now elaborate these concepts in Euclidean n space. For this paragraph let x and y be the n dimensional vectors of the centered data. Assume that each has Euclidean length one, $\|x\| = \|y\| = 1$. Consider the vector $x + y$ in n space; the further this vector is from the origin (which 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. Now 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. To state this another way, for $x - y$ the surface of the centered n-dimensional ball of radius 2, represents perfect negative correlation, so that $\|x - y\|$ large means distance from perfect positive correlation is large. For perfect negative correlation, cos(x, y) =-1, and $\|x - y\|$ = 0, so that the distance from the ball of radius 2 is a maximum.

Summary

condition	effect	cos(x,y)	distance from perfect correlation negative (-1)	positive (+1)
$x \perp y$	$\|x + y\| = \|x - y\|$	0	they are the same	
$x = y$	$\|x + y\| = 2, \|x - y\| = 0$	1	max	min
$x = -y$	$\|x + y\| = 0, \|x - y\| = 2$	-1	min	max

Another way to say this in terms of parameters, is that there is positive correlation when Var(X+Y) > Var(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 standarized normal random variables with CC r . Note that for Z_1 and Z_2 $E(Z_1Z_2) = r = [V(Z_1+Z_2) - V(Z_1-Z_2)] / 4$. Then $V(Z_1+Z_2)$ equals distance from perfect negative correlation and it is a linear function of r , $2+2r$. For r =-1 this distance is zero but for r =+1, this distance is 4. Now $V(Z_1-Z_2)$ is distance from perfect positive correlation and it is $2-2r$. For r =-1 this distance is 4 but for r =+1, this distance is 0. Note that these distances are monotonic functions of r and in this case, for Pearson's CC, the overall correlation $V(Z_1+Z_2) - V(Z_1-Z_2)$ combines to equal $4r$. However, for the other CC this combining of the distance measures does not simplify. Also note that in the case of Pearson's CC

$\frac{1}{2}\ln\frac{V(Z_1+Z_2)}{V(Z_1-Z_2)}=\frac{1}{2}\ln\frac{1+r}{1-r}=\tanh^{-1}r\ =\ln\frac{\sqrt{V(Z_1+Z_2)}}{\sqrt{V(Z_1-Z_2)}}$ which is the Fisher normalizing transformation.

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. See exercise 5 for an example which derives a confidence interval for population CC r based on this ratio definition.

We are now ready to define Pearson's r and a second CC based on absolute values. Let SS_X stand for a corrected sum of squares and SA_X stand for the sum of absolute values about the mean; i.e, $SA_x=\sum|x_i-\bar{x}|$.

Definition 1: Pearson's r

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

= {(standarized distance from perfect negative correlation) -
(standarized distance from perfect positive correlation) } divided by a constant.

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

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

where the denominator is 2 because $\sum\left|\frac{x_i^*}{SA_x}\right|+\sum\left|\frac{y_i^*}{SA_y}\right|=2$.

Note that the same heuristic motivation for Pearson's r holds for this absolute value CC. We are now in a position to define the first NPCC which is based on absolute values. In the same way that Spearman's CC is motivated from Pearson's r by using direct substitution of ranks, so is this new CC obtained from Definition 2 by substitution of ranks. It should be noted, however, that this author obtained the CC in Definition 3 first and determined r_{av} from it.

First note that $(x_i-\bar{x})+(y_i-\bar{y})=x_i+y_i-(\bar{x}+\bar{y})$. Replacing the data by their ranks and ordering the bivariate data by the x data, the ranks are $(i,p_i), i=1,2,\cdots,n$. Thus p_i equals the rank of the y_i for the x_i with rank i. The means of the ranked data are $(n+1)/2$ so that $\bar{x}+\bar{y}$ becomes $n+1$. The ranks p_i are distinct as 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 and put into the denominator. We have

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

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

Definition 3: Spearman's modified footrule correlation coefficient.

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

The square bracket $[*]$ signifies the greatest integer function. The attempt by Spearman (1906) to make an absolute value rank CC is also documented in Kendall and Gibbons(1993). He was trying to make a computationally simple CC and based it onone summation. The idea in this book is that all correlations should be a difference of two functions that measure "distance" from positive and negative correlation which contrasts to Kendall's method in Chapter 2 in Kendall and Gibbons (1993). There, Kendall advances the idea that some type of inner product should be used to define all CC's. The above two absolute value CC's cannot be defined using Kendall's inner product concept. This difference of two functions gives the necessary symmetry to a CC. In Definition 3, the denominator $\left[n^2/2\right]$ comes from the absolute value of the numerator which occurs when $p_i = i$, correlation +1, and when $p_i = n+1-i$, correlation -1. Note again that the same heuristic motivation applies. We now give the definition of Spearman's CC but based on Definition 1.

Definition 4: Spearman's correlation coefficient.

$$r_s(x,y) = \left(\sum(n+1-p_i-i)^2 - \sum(p_i-i)^2\right)/\left((n(n^2-1)/3\right)$$
$$= 1-\left(6\sum(p_i-i)^2\right)/\left(n(n^2-1)\right).$$

The algebra or linear restriction that holds for r_S to simplify the formula does not hold for r_{mf}. Two more CC's are to be defined, Kendall's, for which a linear restriction simplifies 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 CC's symmetric about zero for the case when x and y are independent, the null case. It will also be shown that r_{mf} can be computed from the quantities defined for 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

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

Recall that the data are assumed ordered by the x data and for the i^{th} ranked element of x, the rank of the corresponding y data is p_i. For Kendall's CC, 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. 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 5: Kendall's r_k correlation coefficient.

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

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

$$= \left(4n_c/(n(n-1))\right) - 1 = 1 - \left(4n_d/(n(n-1))\right)$$

since $n_c + n_d = \binom{n}{2}$

where the n choose 2 is one of the standard combination symbols. This summation of n_c and n_d will be shown in the next section to be n choose 2 using a 0-1 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 $d_i^- = \sum_{j=1}^{i} I(n+1-p_j > i)$ be a measure that is large if distance from negative correlation is great.

Definition 6: The Greatest Deviation correlation coefficient, r_{gd}.

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

where $\left\lfloor \frac{n}{2} \right\rfloor$ is the maximun value of the difference in the numerator.

The reader should now do exercises 7 and 8. Now that all of the correlation coefficients have been defined, the next section will give some examples by using a computational aid that augments a plot of the data with a 0-1 matrix.

Section 2: COMPUTATIONS USING THE GRAPH

The data in rank form are $(i, p_i), i = 1, 2, \cdots, n$ and now let $\underline{e}$ = (1,2,....,n) and $\underline{p}$= $(p_1, p_2, \cdots, p_n)$ be the data in vector form. The graph of the ranked data will have $\underline{e}$ plotted on the horizontal axis and $\underline{p}$ plotted on the vertical axis. The data that we shall use is the YMCA basketball data that was used in illustrating the Greatest Deviation CC (Gideon and Hollister, 1987). This data occurred as ranks and it will now be used to calculate all four of the NPCC that have been defined. The $\underline{e}$ will contain 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 $\underline{p}$ $= (14,11,16,2,12,13,7,9,10,3,8,1,15,6,4,5)$. Note that in general the teams with the best won-lost records had the lower sportsmanship ratings. The CC's will put a measure on the relationship between winning and sportsmanship.

The graph of the data appears in Figure 1 in which the *'s indicate the plot points of $(i, p_i), i = 1, 2, \cdots, n$. However, the cartesian product of $\underline{e}$ x $\underline{e}$ on the graph is filled in with 0's being above each of the plotted points and 1's being below. The combination of these *'s, 0's, and 1's will be used to calculate all four NPCC's. The counting technique will be easily seen to be related to the formula's of the NPCC's given in their definitions. To facilitate the method, the two diagonal lines are drawn in; the slope one line, sl^{+1}, $(i, i), i = 1, 2, \cdots, n$ and the slope minus one line, sl^{-1}, $(i, n+1-i), i = 1, 2, \cdots, n$.

Immediately below the graph are two rows which give the values necessary to calculate the Spearman and Absolute Value CC's. 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 Defintions 3 and 4. The sum of the absolute values of these two rows are given just to the right of them, followed by the sum of squares of them.

To the right of the graph are two columns that give the individual concordances and discordances in Kendall's tau as given in Definition 5. A 0 will appear in a column to the right of the * if and only if the rank of that column (the *) is in discordance (less than) and a 1 will appear in a column to the right of the * if and only if the rank of that column is in concordance (greater than). So to obtain the discordances count all the 0's to the right of the * in each column and to obtain the concordances count all the 1's to the right of each * in each column. These results appear in the two columns to the right of the graph. The sum of the two columns, the total numbers of con- and discordances, are given below the line. Note that the ordering within the two columns is not the standard order that is defined as the way to calculate 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 the definition of the Greatest Deviation CC in Definition 6. For each element in the d_i^- column count all the 0's on and to the left of 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.

Figure 1

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

YMCA basketball data: correlation computations

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

For example, let i = 7 in the d_i^+ column, $d_7^+ = \sum_{j=1}^{7} I(p_j > 7) = 5$, since p_1, p_2, p_3, p_5, p_6 are greater than 7. Now i = 7 corresponds to 7 on the vertical axis and counting from the 7 on the vertical axis across to the sl^{+1} diagonal there are 5 one's corresponding to the five p_i's that put one's in row 7 in columns 1,2,3,5,6. For i = 7 in the d_i^- column, $d_7^- = \sum_{j=1}^{7} I(p_j < 17 - 7 = 10) = 2$ since 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 0's 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. These sums will be shown that they can be used to compute r_{mf}. For now note that twice 53 is 106 and twice 28 is 56, the numbers needed for r_{mf}.

From the above statistics which were calculated from Figure 1, the differences in the numerators of the four CC's can be obtained and the denominators are

$$\left[\frac{n^2}{2}\right]=128,\quad n(n^2-1)/3=1360,\quad \binom{n}{2}=120,\quad \left[\frac{n}{2}\right]=8.$$

$$r_{mf}=\frac{56-106}{128}=\frac{-25}{64}=-.3906,\quad r_s=\frac{348-1012}{1360}=\frac{-83}{170}=-.4882$$

$$r_k=\frac{38-82}{120}=\frac{-11}{30}=-.3667,\quad r_{gd}=\frac{3-6}{8}=\frac{-3}{8}=-.3750.$$

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), a linear restriction, but that this does not occur for r_{gd} and r_{mf} (r_{gd}: 6 + 3 = 9 > 8, and r_{mf}: 56 + 106 = 162 > 128).

SPECIAL FORM FOR CALULATION OF r_{gd}

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

Table 1: Calculation of the Greatest Deviation CC

i	p_i	n+1-p_i	d_i^+	d_i^-
1	14	3	1	1
2	11	6	2	2
3	16	1	3	1
4	2	15	3	2
5	12	5	4	2
6	13	4	5	1
7	7	10	5	2
8	9	8	6	2
9	10	7	6	2
10	3	14	5	2
11	8	9	4	2
12	1	16	3	3
13	15	2	3	3

14	6	11	2	2
15	4	13	1	1
16	5	12	0	0
			max = 6	max = 3

THREE LEMMAS

We now show the relationship between the statistics used in r_{gd} and r_{mf}.

Lemma 1: $2\sum d_i^+ = \sum|p_i - i|$ and $2\sum d_i^- = \sum|n+1-p_i - i|$.

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 1's on or above the sl^{+1} line. But, $d_i^+ = \sum_{j=1}^{i} I(p_j > i)$ counts all the 1's 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.

Lemma 2: In this lemma the i in row i refers to the vertical axis which are ranks; e.g. row 1 corresponds to the bottom row of the 0-1 graph matrix. The number of 1's on or to the right of the sl^{-1} line in row i-1 equals the number of 0's on or to the left of sl^{-1} in row i, i=2,3,...,n. The number of 0's on or to the right of the sl^{+1} line in row i equals the number of 1's on or to the left of the sl^{+1} line in row i-1, i=2,3,...,n.
Proof: Left to the reader, it can be checked in Figure 1.
The symmetry displayed in this Lemma shows that the r_{gd} CC could have been equivalently defined in a right-handed fashion.

Lemma 3: For Kendall's CC, $n_c + n_d = \binom{n}{2}$
Proof: The plotted values (*'s) divide the 0-1 matrix into left and right parts with a total of $n^2 - n$ 0's and 1's. By the symmetry of the graph the number of 0's and 1's to the left of the *'s must equal the number to the right. Thus, $n_c + n_d = \frac{n^2 - n}{2} = \binom{n}{2}$. Even further, the number of 0's to the left (38 in Figure 1) equals the number of 1's to the right and the number of 1's to the left (82 in Figure 1) equals the number of 0's to the right.

Section 3: WHICH CC'S ARE OUTLIER RESISTANT?

In this section two examples are given to illustrate that the four NPCC's 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). These examples 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 $e = (1, 2, \cdots, n)$ and $p = (p_1, p_2, \cdots, p_n)$. The calculation of the CC's are left as exercise 4. The values of the NPCC's for n = 10 on e and $p = (5, 4, 3, 2, 1, 10, 9, 8, 7, 6)$ are

$r_{mf} = {}^{26}\!/_{50} = .5200,\ r_s = {}^{17}\!/_{33} = .5152,\ r_k = {}^{1}\!/_{9} = .1111,\ r_{gd} = {}^{3}\!/_{5} = .6000$.

The values of the CC's on e again but with $p = (10, 2, 3, 4, 5, 6, 7, 8, 9, 1)$ are

$r_{mf} = {}^{14}\!/_{50} = .2800,\ r_s = {}^{6}\!/_{330} = .0182,\ r_k = {}^{11}\!/_{45} = .2444,\ r_{gd} = {}^{3}\!/_{5} = .6000$.

It is known that for the bivariate normal distribution, the NPCC estimate a function of the correlation parameter r that is less than r. Thus, when the NPCC are lot greater than r_s, it is clear that there are "strange" observations in the data. In the "strange" data of these two examples, clearly r_{gd} and r_{mf} give the largest indication of a positive relationship and hence, may be the most resistant to outliers.

Section 4: PROBABILITES AND ASYMPTOTICS FOR THE RANK CORRELATIONS

Some aspects of the rank CC 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 distinquish between two sound tones. Spearman used this example to illustrate his footrule CC. The data was for eleven students of psychology and their ability in addition and pitch discrimination was ranked independently by Spearman (sound) and a second person for addition. The data are ordered by the addition variable.

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)

Spearmans's Footrule CC was $r_f = 1 - \dfrac{6\sum_{p_i > i}(p_i - i)}{n^2 - 1} = 1 - \dfrac{6(8.5)}{120} = 0.57$. He compared this number to "probable error" (that is derived in his paper) of 0.13 and concluded that since 0.57/0.13 = 4.38, "the faculty of adding numbers and that of discrimination pitch is just about large enough to be beyond all reasonable suspicion of mere chance coincidence." Spearman used the notation "g" for this CC, and "g" back then, and still today (see

Herrnstein and Murray ,1994) was used to denote a measure of general intelligence or now called "cognitive ability" in Herrnstein and Murray (1994). The four nonparametric CC and their corresponding probability values will now be computed for this data. With regard to 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". Thus, Spearman wanted a robust CC for his data.

This example will also illustrate the definition of a rank CC when tied values are present. In Table ---- , the calculations of the four rank CC are done when Person B is assigned rank 4 for sound and Person J is assigned rank 5 for sound. These ranks are circled stars in the rank-graph. When ties are broken in the reverse direction, the effect on the calulations of the four CC are shown by blocking. Note that r_{gd} is the only CC without a change. Each CC can be defined uniquely by averaging the values of its two possible values.

The effect of interchanging the sound ranks for Persons B and J was to produce sets of ranks that in one case were broken to most favor positive correlation, Person B gets sound rank 4 since its addition rank is 4, and in the other case to most favor negative correlation, Person J with addition rank 5 gets the lower sound rank 4. In Table ---, r_{gd} remains at 0.6000 but r_{mf} is defined to be (0.7333 + 0.7000)/2 = 0.7167. A general definition is now given.

DEFINITION: Values of rank CC when ties are present

Let (x,y) be a set of data, and let (I^+,P^+) be the corresponding ranks which are assigned in such a way that most favor positive correlation, and let (I^-,P^-) the corresponding ranks assigned in a way to most favor negative correlation. The assignment of ranks must be done within the blocks of tied values. "I" will become $(1,2,\cdots.n)$ and P will be a permutation of this I. Then a rank CC r_* is defined uniquely on P^+ and P^- and its value is

$r_*(x,y) = (r_*(P^+) + r_*(P^-))/2$.

We abbreviate $r(P^+)$ and $r(P^-)$ 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},

$$r_{gd} = \frac{r^+ + r^-}{2} = \frac{1/4 + (-1/4)}{2} = 0$$

.

We now return to the level of significance for the Spearman example. These are obtained from Neave (1978) for r_k and r_s, from Gideon and Hollister (1987) for r_{gd}, and from an unpublished table for r_{mf}. For r_{mf} the table given in this text only goes up to sample size 10 and it will be shown that the limiting distibution is quite good for this example at n = 11. 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, in Gideon, Pyke, and Prentice (1989) for r_{gd}, and from this text for r_{mf}. The asymptotic null distributions (r = 0) of the four CC are given first; $\sqrt{n-1}r_s : N(0,1)$;

$\sqrt{n-1}r_k : N(0,4/9)$; $\sqrt{n}r_{gd} : N(0,1)$; $\sqrt{n-1}r_{mf} : N(0,2/3)$. 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 are treated for the most correlation case. First, from tables, $0.001 \le P(r_s \ge 0.7636) \le 0.005$; $0.01 \le P(r_k \ge 0.5636) \le 0.025$; $0.01 \le P(r_{gd} \ge 0.6000) \le 0.05$; $P(r_{mf} \ge 11/15 = 0.7333) = 0.0013$ and $P(r_{mf} \ge 7/10 = 0.7000) = 0.0024$. Thus, all of the CC are significant with r_S and r_{mf} being the most significant. These results are now compared to the asymptotic approximations (let Z be N(0,1)).

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

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

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

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

.

All of these approximate results are reasonably good. All four correlations support Spearman's conclusion that his "footrule" CC gave; i.e., "that our correlation between 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". 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 $\sqrt{V(r_{mf})} = \sqrt{\frac{2(11^2+3)}{3(10)(11^2-1)}} = \sqrt{0.0689} = 0.2625$; 0.7333/0.2625 = 2.7937. By Spearman's rules of "satisfactory demonstration" that this ratio be at least 5, r_{mf}, the completion of his footrule CC is not nearly as significant as his footrule CC. 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.