Location and Scale Estimation with Correlation Coefficients

Rudy Gideon and Adele Marie Rothan, CSJ

University of Montana
Missoula, MT 59812

College of St. Catherine
St. Paul, MN 55105

1. Abstract

This paper, third in a series on the correlation estimation system (CES), shows how to use any correlation coefficient to produce an estimate of location and scale. Since the normal distribution is so widely used, the method is illustrated using this distribution. Analyzers of normal data are advised to graph a quantile plot to check on the normality assumption before performing their data analysis; (Looney and Gulledge, 1985) shows how to use Pearson's r_p as a test of normality. This paper shows that any correlation coefficient can be used to fit a simple linear regression line to a graph and then the slope and intercept are estimates of standard deviation and location. Because a robust correlation will produce robust estimates, this CES can be recommended as a tool for everyday data analysis. Tables of mean square error for simulations indicate that the median with this method using a robust correlation coefficient appears to be nearly as efficient as the mean with good data and much better if there are a few errant data points. Hypothesis testing and confidence intervals are illustrated for the scale parameter.

Key words: simple linear regression, robust estimates, hypothesis testing, confidence intervals

This work depends in part on earlier unpublished work of Gideon and others and is available on the web site: www.math.umt.edu/gideon. Some of the references will refer to papers posted at this web site.

2. Introduction

As in (Gideon, 1992) three correlation coefficients (CC) are used: Pearson's r_p, Kendall's t, and Greatest Deviation Correlation Coefficient (*GDCC* or r_{gd}), (Gideon and Hollister, 1987). The starting point for each estimation technique is exactly the same. The CCs chosen illustrate existing techniques: Pearson's, classical statistics; *GDCC*, robust methods; Kendall's t, a well-known nonparametric (NP) CC. A problem in (Randles and Wolfe, 1979, p. 12, problem 1.2.14) indicates how to estimate location and scale from order statistics. This method is reviewed and then its connection to Pearson's r_p is made for data from a normal distribution. Note, however, that the method is general for any distribution that can be standardized.

Let $Y = m + s\ Z$ where Z is normal with mean 0 and standard deviation 1, denoted by $Z \sim N(0,1)$; then $Y \sim N(m, s)$. Then for the order statistics $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$, $Y_{(i)} = m + s\ Z_{(i)}$ and $E(Y_{(i)}) = m + sE(Z_{(i)})$. Let $k_i = E(Z_{(i)}),\ i = 1,2,\cdots,n$. From the symmetry of the standard normal, note that $\sum k_i = 0$. (Randles and Wolfe, 1979) next

-

define $D(m, s) = \sum_{i=1}^{n} (Y_{(i)} - (m + s\, k_i))^2$. The estimators $\hat{m}$ and $\hat{s}$ that are found to minimize D are unbiased for m and s, respectively.

This solution is next related to Pearson's r_p. In general, let *k* be the vector of the expected values of the order statistics of *Z*, and let y^o be the order statistics of a sample from *Y*; *i.e.*, y^o represents the order statistics $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$. If the following equation is solved for *s* using any *r*, then *s* estimates s:

$$r(k, y^o - sk) = 0 . \qquad (1)$$

Using Pearson's r_p for the *r* let the uncentered residuals $y^o - sk$ be denoted by res and compute the mean of res after *s* has been determined. This mean estimates m ; in fact, these latter two estimates are identical to the ones coming from $D(m, s)$. From Publication 2, *Correlation in Simple Linear Regression* (see www.math.umt.edu/gideon), with *x* = *k* and $y = y^o$, the regression equation (1) of that paper becomes the above (1), called the scale form of the regression equation. The solution is $s = \dfrac{\sum k_i y_{(i)}}{\sum k_i^2}$ and

$\text{mean}(res) = \bar{y} - s\dfrac{\sum k_i}{n} = \bar{y}$. Note that the usual estimate of the mean is obtained. For the estimate of s the statistic *s* is unbiased because

$$E(s) = \frac{\sum k_i E(Y_{(i)})}{\sum k_i^2} = \frac{\sum k_i (m + s\, k_i)}{\sum k_i^2} = \frac{m \sum k_i + s \sum k_i^2}{\sum k_i^2} = s .$$

-

The use of equation (1) with Pearson's r_p as a scale estimation technique is now related to two existing scale estimators. Motivated from (Downton, 1966), let

$$k = \frac{6}{(n+1)\sqrt{p}} \left\{ \left[\begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} - \frac{n+1}{2} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right] \right\} \qquad (2)$$

The solution for *s* in equation (1) with this *k* is related to both Gini's mean difference (Randles and Wolfe, 1979) or (Hettmansperger, 1984) and a method of Downton (1966). Gini's mean difference estimate of scale is $D(y) = \frac{1}{\binom{n}{2}} \sum_{i<j} \left| y_{(i)} - y_{(j)} \right|$, and Downton's estimate of scale for the normal distribution is $s_{dt} = \frac{\sqrt{p}}{\binom{n}{2}} \sum_{i=1}^{n} (i - \frac{n+1}{2}) y_{(i)}$.

It can be shown that $s_{dt} = \frac{\sqrt{p}}{2} D(y)$ so that Gini and Downton are essentially the same, and both can be obtained from equation (1) with the *k* given in (2). Thus, with today's computers and statistical packages, all of the above estimates of scale can be obtained easily from the regression setting (equation (1)) with the ordered data *y* and an appropriate *k*.

(D'Agostino, 1971 & 1973) used Downton's estimate of scale divided by the classical least squares estimate of s to perform a test of normality. The estimate of s from equation (1) with $k_i = E(Z_{(i)})$, $i = 1,2,\cdots,n$ could also be used in the D'Agostino normal test of fit with this *s* replacing the classical estimate.

-

An interpretation of the usual SD = $\sqrt{\dfrac{\sum(y_i - \bar{y})^2}{(n-1)}}$ as the slope of a straight line is next used as a transition to a more geometrical view of scale estimates. Again consider the data ordered $y_1 < y_2 < \cdots < y_n$ and let constant $c = \sqrt{12/(n(n+1))}$. For the horizontal axis points let $h = -\dfrac{n-1}{2}, -\dfrac{n-3}{2}, -\dfrac{n-5}{2} \cdots 0, 1, \cdots \dfrac{n-5}{2}, \dfrac{n-3}{2}, \dfrac{n-1}{2}$. For simplification only the case n odd is used so that h consists of n integers centered at zero (the even case only requires a change in notation). Now consider the set of points (ch, y^o) where the superscript indicates the ordered vector of data points. A line whose slope is SD is chosen below to go through these points. Let a horizontal line be drawn at the mean of the data, $\bar{y}$, on the vertical axis. The distance of each order statistic from the $\bar{y}$ line measures its departure from that line; the SD is an overall measure of departure from the horizontal. A plot of (ch, y^o) roughly creates an angle q from the horizontal with $\tan q = SD$.

With this motivation, a straight line with slope $b = SD$ and intercept $\bar{y}$ is now constructed through the points (ch, y^o). This requires that the line be chosen so that the points bch have the same cumulative squared distance from the $\bar{y}$ line. Hence b is chosen so that

$\sum(y_i - \bar{y})^2 = \sum_{h=-\frac{n-1}{2}}^{\frac{n-1}{2}} (bhc)^2$. The line with slope b defines an angle q with the horizontal.

Because $\sum_{h=-\frac{n-1}{2}}^{\frac{n-1}{2}} h^2 = 2\sum_{h=1}^{\frac{n-1}{2}} h^2 = \dfrac{n(n-1)(n+1)}{12}$ and using the definition of c, the b that satisfies the above equation is $b = SD.$ The figure at the end of this paper shows this line for a normal random sample, n = 25 with mean 10 and theoretical standard deviation 7.

-

For this data $\bar{y}$ = 10.56 and SD = 7.39 and so the vertical values are $10.56 + 7.39 * (ch)$. The slope b is the same as SD and represents the variation in the data. A steeper slope (or a larger q) implies more variation and a 0 slope (or q = 0) indicates no variation. The range of the plot on the horizontal axis, regardless of the data set, is roughly between $\pm\sqrt{3}$ for all n and has equidistant points.

The plotted points (ch, y^o) can be used to illustrate the CES way of estimating s . The scale regression equation (1) is solved for s_p, the slope, using Pearson's r_p . The equation is $r_p(ch,\ y^o - s_p hc) = 0$. It is straightforward to obtain s_p = $\sqrt{\frac{12}{n(n+1)}}\frac{1}{n-1}\sum_{h=-\frac{n-1}{2}}^{\frac{n-1}{2}} hy_{(h)}$. For the data used on the Figure $s_p = 7.24$ and for GDCC the scale estimate was 6.80. After a few computer runs it was clear that s_p and SD or b have about a 99% Pearson correlation. However, s_p is slightly biased. An adjustment to the constant c would make s_p unbiased, which is what Downton's estimate does. His constant is $\frac{6}{(n+1)\sqrt{p}}$ as compared to c. The ratio of Downton's constant to c is $\sqrt{3(n+1)/pn}$. Asssuming $(n+1)/n \cong 1$ this is about 0.9772. Downton constructed his constant so that the r_p scale regression solution is unbiased. Downton used linear combinations of order statistics as his approach rather than using correlation as is done here.

In addition to using CCs in tests of fit (distribution), the CC can be used to estimate location and scale as in the example above. Equation (1) can be solved with any

-

correlation coefficient, r. The next section continues the demonstration of the method using *GDCC* and Kendall’s τ. After obtaining s, either the mean or median of the uncentered residuals is used to obtain a location estimate of the y-data.

3. Interpreting Equation (1)

When r_{gd} is used for r in equation (1) the solution s must be found numerically as no closed form solution is known, but for Kendall's t, the equation $t\,(k, y^o - sk) = 0$ is satisfied by $s = median\left(\frac{y_{(j)} - y_{(i)}}{k_j - k_i}\right)$(see web site Publication 2). Because of the discrete nature of both of the NPCCs, a range of solutions is possible, so a unique s is defined by letting $s = (s_l + s_u)/2$ where, $s_l = \sup\{s : r(k, y^o - sk) > 0\}$ and $s_u = \inf\{s : r(k, y^o - sk) < 0\}$. This averaging obtains a unique solution for either r = r_{gd} or r = t, or for that matter any NPCC.

Note that the left-hand side of equation (1) is a function of $y, s = s(y)$. The function $s(y)$ has the following form for each of the three CCs considered:

1. for Pearson's r_p, $s(y)$ is a continuous function and (1) has a closed form solution;
2. for *GDCC*, $s(y)$ is a step function based on a NPCC and (1) has only a numerical solution;
3. for Kendall’s t, $s(y)$ is a step function based on a NPCC and (1) has closed form solution.

-

3. Standard Properties of the Scale Estimator, $s(Y)$

The function s(*Y*) is next shown to be location invariant, scale equivariant, and for symmetric distributions, $s(Y) = s(-Y)$; *i.e.* it is even. Because CCs are location invariant, $s(Y + h*1) = s(Y)$, where h is any constant and 1 is a n-vector of all 1s and so $s(Y)$ is location invariant. Keep in mind that all data are ordered even though the "superscript ^{0}" notation is not always used. (Rousseeuw and Leroy, 1987) use the term equivariant for statistics that transform properly. Note $s(Y)$ is scale equivariant; i.e., if $h > 0$ is a constant and $X = hY$ is a scale change, then it is easy to show $s(hY) = hs(Y)$. Because $r(k, X - s(X)k) = 0$ and CCs are scale invariant, seeing that $r(k, X - hs(Y)k) = r(k, hY - hs(Y)k) = r(k, Y - s(Y)k) = 0$, verifies that $hs(Y)$ is $s(X)$, that is, $s(hY) = hs(Y)$. The evenness of $s(Y)$ for a NPCC requires a lemma.

Lemma 1: Given a NPCC in equation (1) and a symmetric distribution about 0, $s(Y) = s(-Y)$.

Proof: Since the distribution is symmetric about 0, $k_{n+1-i} = -k_i$, $i = 1,2,\cdots,n$ and for the vector k, $(-k)^o = k^o = k$. It is also true that
$$(-y)^o = \begin{pmatrix} -y_{(1)} \\ -y_{(2)} \\ \vdots \\ -y_{(n-1)} \\ -y_{(n)} \end{pmatrix}^o = \begin{pmatrix} -y_{(n)} \\ -y_{(n-1)} \\ \vdots \\ -y_{(2)} \\ -y_{(1)} \end{pmatrix}.$$

-

In equation (1), $0 = r(k,(-y)^0 - s(-y)k) = r((-k)^0,(-y)^0 - s(-y)*(-k)^0)$, and $(-k)^o$ and $(-y)^o$ are ordered min to max. Without the superscript ^{0}, they still correspond but are now ordered max to min. So in equation (1),

$$0 = r((-k),(-y)-s(-y*(-k))) = r(-k,-(y-s(-y)*k)) = r(k,y-s(-y)*k);$$

the right-most term being equal to zero shows that $s(Y) = s(-Y)$. ♦

This lemma is easily demonstrated for particular cases on a computer.

5. Motivation and Standard Properties of the CES Location Estimator of *Y*

Because CCs are the estimation tools, the location estimator of *Y*, say *l*(*Y*), is motivated through regression; the result for Pearson's r_p is the classical mean of the data, whereas for Kendall's τ and *GDCC* it is the median. To motivate these results, first consider data from two independent random variables *X* and *Y* with sample sizes *m* and *n*, respectively. The location difference between the two samples is studied via regression. On a coordinate plane, let the *x*-data be plotted as $(0,x_i)$ for $1 \le i \le m$ and the *y*-data as $(1,y_i)$ for $1 \le i \le n$. If there is no difference in the *X* and *Y* locations, then a line connecting the center of the *x*-data to the center of the *y*-data should be nearly parallel to the horizontal axis. To estimate any possible location difference, a regression line is fit with a coded variable and the (*x*, *y)* data. Let the column vector *c* of dimension *m* + *n* be given by *m* 0s followed by *n* 1s and the *m* + *n* dimension vector *v* be $(x_1, x_2, \cdots, x_m, y_1, y_2, \cdots, y_n)'$. Treat *c* as the regressor variable and *v* as the response variable. Then the CC regression equation is $r(c, v - lc) = 0$ where l is a location statistic. It is straightforward to solve this equation

-

with Pearson's r_p to obtain $l = \bar{y} - \bar{x}$. Thus, the slope is $\bar{y} - \bar{x}$ and $\bar{x} + slope = \bar{y}$. For the one-sample problem, let all of the *x*- data be zero; then the estimate of the location of the *y*-data is the slope $\bar{y}$ since $\bar{x}$ is zero.

To solve $t\,(c, v - lc) = 0$ it is necessary to work with the elementary slopes of *c* and *v*, $\frac{v_j - v_i}{c_j - c_i}$, where they are finite, that is, where $c_j - c_i = \pm 1$. This results in *l* being the median of the *mn* elementary differences $y_j - x_i$. For the one-sample case, all the *x*'s are zero, so $l = median(y_j)$. As discussed in (Gideon and Rummel, 1992) if the *x*-data are all zeros and have the same dimension as *Y*, namely *n*, and in addition if the tied value method (Gideon and Hollister, 1987) is used in the calculation of the NPCCs, then for both τ and *GDCC* the median is obtained as the solution to the regression equation, $r(c, v - lc) = 0$. This has not been proven for *GDCC*, but only demonstrated via extensive computer simulations. This computer work and analysis shows that both the one- and two-sample problems posed in a regression setting can be performed for NPCCs as has been done for the least squares (Pearson's r_p) regression method. The implication is that a fertile field of research awaits generalization to analysis of variance via regression with NPCCs.

Because the location estimator for Pearson's r_p is the usual $\bar{y}$, it is obviously an odd translation statistic; i.e. a location statistic. For the other two CCs, $l(y) = median(y^o - sk)$ where $s = s_t$ or $s = s_{gd}$ is the solution of $r(k, y - sk) = 0$ and *r* is τ or *GD*CC respectively.

-

Lemma 2: For NPCCs τ and *GDCC* and for a symmetric distribution,

$l(y) = median(y^o - sk)$ is an odd translation statistic.

Proof: $l(-y) = median\left((-y)^o - sk\right) = median\left((-y)^o - s(-k)^o\right)$

$$= median((-y) - s(-k)) = -median(y - sk) = -l(y)$$

For translation, with constant *h*,

$l(y+h) = median((y+h)^o - sk) = h + median(y^o - sk) = h + l(y)$.

Therefore, the location estimator with NPCCs also has the properties of a location statistic.

♦

Because there is a closed form solution of the scale regression equation (1) using Kendall's τ, it is possible to make a closer examination of its scale and location estimates.

Let the elementary slopes be $l_{ji} = \dfrac{Y_{(j)} - Y_{(i)}}{k_j - k_i}$, for $1 \le i < j \le n$ where $k_i = E(Z_{(i)})$. Now

$E(l_{ji}) = \dfrac{E(Y_{(j)}) - E(Y_{(i)})}{k_j - k_i} = \dfrac{(m + sk_j) - (m + sk_i)}{k_j - k_i} = s$. Each l_{ji} can be considered a

random observation from a population with mean s; therefore, $E(mean(l_{ji})) = s$.

However, to be unbiased, the scale estimator, $s_t\,(y) = median(l_{ji})$, depends on the symmetry of the distribution of the correlated l_{ji}. The quantity s_t is either the mean of the two central order statistics or the middle order statistic of the l_{ji} whose expectation is, in any case, s. Table 2 shows that s_t appears to have a slight positive bias in estimating SD. If $res_i = y_{(i)} - s_t k_i,\ i = 1,2,\cdots,n$ and $E(s_t\,(y)) = s^+ > s$, then

-

$E(res_i) = E(y_{(i)} - s_t k_i) = (m + s\ k_i) - s^{+} k_i \ = m + (s\ - s^{+})k_i$. Because each residual, res_i, has expectation possibly slightly less than m for $k_i > 0$, but slightly greater for $k_i < 0$, the expectation of the median of the residuals may be approximately m. In the simulation results, the positive bias in the estimation of scale is apparent; but no bias seems to appear in the estimation of location.

The "equal in distribution" technique described in (Randles and Wolfe, 1979, Section 1.3) can be used to show that $s_t\,(y)$ and $l_t\,(y)$ are uncorrelated statistics. Of course, for the normal distribution, the classical estimate of s and the sample mean are independent. Whether or not this independence result is true for the estimators based on other CCs is unknown.

This section concludes with a proof that the location estimator, $l_t\,(y)$, is symmetrically unbiased. In CES, it is necessary to first estimate the scale and then the location.

Assume $Y^* - m \overset{d}{=} m - Y^*$; i.e. Y^* is symmetric about m. Then without loss of generality, $Y = Y^* - m$ is symmetric about zero. The distribution function $F(y)$ is

$F(y) = P(Y \le y) = P(Z \le \dfrac{y-m}{s})$. Because m = 0, $Y_{(i)} = s\ Z_{(i)},\ i = 1,\ 2, \cdots, n$. The estimate of the standard deviation with Kendall's τ, s_t, is

$s_t \ = \underset{i<j}{median}\left(\dfrac{y_{(j)} - y_{(i)}}{k_j - k_i}\right)$ where $k_i \ = E(Z_{(i)})$. Because

-

$E\left(\dfrac{Y_{(j)}-Y_{(i)}}{k_j-k_i}\right)=s\left(\dfrac{E(Z_{(j)})-E(Z_{(i)})}{k_j-k_i}\right)=s$, it is expected that s_t would be a reasonably good estimate of the standard deviation s.

Earlier it was shown in Lemma 1 that $s(Y) = s(-Y)$, but it is constructive to show this again specifically for Kendall's τ; that is, $s_t(y) = s_t(-y)$. Let $X = -Y$ or for a random sample $x_i = -y_i$. Then for order statistics, $x_{(i)} = -y_{(n+1-i)}$, $i = 1, 2, \cdots, n$ and

$$s_t(x) = \underset{i<j}{median}\left(\frac{x_{(j)}-x_{(i)}}{k_j-k_i}\right) = median\left(\frac{-y_{(n+1-j)}+y_{(n+1-i)}}{k_j-k_i}\right)$$

Now $k_j = -k_{n+1-j}$ by the symmetry assumption, so

$$s_t(x) = median\left(\frac{y_{(n+1-i)}-y_{(n+1-j)}}{k_{n+1-i}-k_{n+1-j}}\right) = median\left(\frac{y_{(j)}-y_{(i)}}{k_j-k_i}\right) = s_t(y)\ .$$

Lemma 3: Kendall's τ estimate of the median of a symmetric distribution has a symmetric distribution about the true population median (mean); that is, $l_t(-y) = -l_t(y)$.

Proof: The estimate of the population median based on the residuals of the scale estimate is $l_t(y) = median(y_{(j)} - s_t(y)k_j)$. Let, as above, $X = -Y$. Then,

$$\begin{aligned} l_t(-y) = l_t(x) &= median\left(x_{(j)} - s_t(x)k_j\right) \\ &= median\left(-y_{(n+1-j)} - s_t(x)\left(-k_{n+1-j}\right)\right) \\ &= -median\left(y_{(n+1-j)} - s_t(x)\left(k_{n+1-j}\right)\right) \\ &= -median\left(y_{(n+1-j)} - s_t(-y)\left(k_{n+1-j}\right)\right) \\ &= -median\left(y_{(n+1-j)} - s_t(y)\left(k_{n+1-j}\right)\right) \text{ because } s_t(y) = s_t(-y) \end{aligned}$$

-

$$= -median\left(y_{(j)} - s_t(y)\left(k_j\right)\right)$$

$$= -l_t(y)$$

By theorem 1.3.16 in (Randles and Wolfe, 1979, p. 20), since $Y \overset{d}{=} -Y$ and $l_t(-y) = -l_t(y)$, the distribution of $l_t(y)$ is symmetrically distributed about zero. Thus, we can say that $l_t(y)$ is symmetrically unbiased. ♦

6. A Simulation Study of the Scale and Location Estimates of *GDCC* and Kendall's τ

Although the expected values of the order statistics are available, e.g., (Harter and Balakrishnan, 1996) up to *n* = 400 for the Normal, most statistical computer packages do not have them readily available. They can be approximated — see (Gibbons and Chakraborti, 1992, Section 2.6) — and a first approximation is given by $\Phi^{-1}(\frac{i}{n+1}),\ i = 1,2,\cdots,n$ where Φ is the distribution function of a *N*(0, 1) random variable. This approximation seems to work well but, rather than $p_i = \frac{i}{n+1}$, other p_i's are recommended for different sample sizes (David, 1970 and Looney and Gulledge, 1985). In this simulation study on location and scale estimation, *S-Plus* was used since Φ^{-1} and other distribution functions are available and the language lends itself to investigative inquiry (Venables and Ripley, 1994).

The estimation of s and m by the two NPCCs is studied via computer simulation. In Tables 1 and 2, the normal distribution was used with mean 10, standard deviation 7, and sample

-

size n = 25. Kendall's τ and *GDCC* were compared separately to the classical estimators, sample standard deviation (SD) and sample mean. The first part of each table gives the mean estimate of the parameters, m or s; and the second part gives the square root of the sample mean square error of the estimators, $\sqrt{MSE}$, based on the four sets of runs of 250 simulations labeled 1, 2, 3, 4. Table 1 gives the results of *GDCC* compared to SD and the classical median and mean. Note that mean values of s_{gd} are slightly high but that l_{gd} is nearly unbiased. The $\sqrt{MSE}$ of s_{gd} is somewhat higher than SD, but the most interesting aspect is that the $\sqrt{MSE}$ of l_{gd} is lower than the classical median and just barely larger than the $\sqrt{MSE}$ of the sample mean.

Table 1 Comparison of *GD*CC and Classical Estimates for s and m

	$s = 7$		$m = 10$		
Run #	s_{gd}	SD	l_{gd}	Median	Mean
1	7.28	6.97	9.82	9.87	9.82
2	7.28	7.00	10.10	10.06	10.11
3	7.13	6.91	9.88	9.82	9.91
4	7.09	6.83	9.92	9.86	9.91

$\sqrt{\text{Mean Square Error}}$

	$s = 7$		$m = 10$		
Run #	s_{gd}	SD	l_{gd}	Median	Mean

-

1	1.33	0.97	1.36	1.67	1.34
2	1.31	1.03	1.43	1.73	1.38
3	1.43	1.06	1.49	1.88	1.45
4	1.33	1.06	1.33	1.64	1.31

All Runs: $n = 25$, $N(10, 7)$; 250 simulations for each run.

Each entry is the mean of the results of 250 simulations.

These same observations are repeated for Table 2 with Kendall's τ compared to the classical estimators. The $\sqrt{MSE}$ for s_t is lower than that of s_{gd} and only about 17% higher than that of SD.

Table 2 Comparison of Kendall's τ and Classical Estimates for s and m

	$s = 7$			$m = 10$	
Run #	s_t	SD	l_t	Median	Mean
1	7.22	6.92	9.90	9.83	9.90
2	7.14	6.90	9.98	9.99	9.99
3	7.17	6.86	10.05	10.01	10.04
4	7.34	7.09	10.15	10.09	10.15

$\sqrt{\text{Mean Square Error}}$

	$s = 7$			$m = 10$	
Run #	s_t	SD	l_t	Median	Mean

-

1	1.11	0.95	1.42	1.83	1.37
2	1.06	0.95	1.35	1.67	1.31
3	1.15	1.00	1.47	1.78	1.42
4	1.23	0.99	1.42	1.79	1.41

All Runs: n = 25, N(10, 7); 250 simulations for each run.

Each entry is the mean of the results of 250 simulations.

Tables 3 and 4 again compare *GDCC* and Kendall's τ methods to classical methods. Twenty of the random observations are from N(10, 7), but now 5 of the 25 observations can be outliers. On Runs 1 and 2, the five outlier observations are from a N(10, 35) random variable; and on runs 3 and 4, the five outlier observations are from N(17, 35). Thus, runs 1 and 2 have centered outliers while runs 3 and 4 have right-biased outliers.

Table 3 gives results for *GDCC*, and Table 4 for τ. By far, s_{gd} is the best estimator of scale with less bias and smaller $\sqrt{MSE}$; s_t is also far better than SD. For location, all the median methods are much better than the classical mean. The classical median and the median methods l_{gd} and l_t have mean values close to 10, and the values of $\sqrt{MSE}$ are not too different although l_{gd} did have $\sqrt{MSE}$ lower than the classical median in all four sets of simulations. It appears that the NPCC method of scale and location gives good protection against a few errant observations and, at the same time, loses very little if all the data are good. Because of this, the CES procedure should not be ignored.

-

Table 3 Comparison of *GDCC* and Classical Estimates for s and m , Data with Outliers

	s		m		
Run #	s_{gd}	SD	l_{gd}	Median	Mean
1	9.59	15.82	10.06	10.12	9.97
2	9.65	15.60	10.08	10.13	10.35
3	9.59	16.57	10.67	10.52	11.45
4	9.71	16.71	10.55	10.41	11.42

$\sqrt{\text{Mean Square Error}}$

	s		m		
Run #	s_{gd}	SD	l_{gd}	Median	Mean
1	3.13	9.83	2.10	2.27	3.39
2	3.33	9.78	1.91	1.97	3.04
3	3.16	10.55	2.14	2.21	3.68
4	3.20	10.78	2.10	2.13	3.82

Runs 1 and 2: n = 25, 20 of N(10, 7) & 5 of N(10, 35); 250 simulations for each run.

Runs 3 and 4: n = 25, 20 of N(10, 7) & 5 of N(17, 35); 250 simulations for each run.

Each entry is the mean of the results of 250 simulations.

Table 4 Comparison of Kendall's τ and Classical Estimates, Data with Outliers

	s		m		
Run #	s_t	SD	l_t	Median	Mean
1	10.23	16.06	9.91	9.90	10.01
2	10.37	16.44	9.67	9.80	9.36
3	10.29	16.76	10.71	10.41	11.77
4	10.30	16.37	10.51	10.30	11.27

-

$$\sqrt{\text{Mean Square Error}}$$

	s		m		
Run #	s_t	SD	l_t	Median	Mean
1	3.78	10.02	2.05	2.11	3.25
2	3.79	10.34	2.07	2.00	3.53
3	3.78	10.65	2.07	1.97	3.75
4	3.81	10.36	2.16	2.03	3.67

Runs 1 and 2: n = 25, 20 of N(10, 7) & 5 of N(10, 35); 250 simulations for each run.

Runs 3 and 4: n = 25, 20 of N(10, 7) & 5 of N(17, 35); 250 simulations for each run.

Each entry is the mean of the results of 250 simulations.

(Chambers, Cleveland, Kleiner, and Tukey, 1983) gives background on the general use of Quantile-Quantile plots and in Section 6.8 gives the parameters that are estimated by the intercept and slope of a line fit to the data. Thus, for example, if a test of fit is desired for a Gamma random variable and then a CES linear regression is fit, the parameters estimated by this fitted line are given. These authors use a different choice of p_i not $\frac{i}{n+1}$.

7. Hypothesis Testing and Confidence Intervals for s

There is an acute need for a better scale analysis because most scale tests under the normality assumption lead to unreliable results. This section will indicate how such a test is performed utilizing the work of (Gee, 2002) and (Gideon and Rummel, 1992) as well as the earlier sections of this paper. With the speed of computers and their many functional

-

statistical packages, all of the following can be done by anyone wanting to implement the strategy; e.g. critical values can be estimated by simulations. Limited resources have not allowed a full study of the ideas and the sorting out of which CCs might be most useful in hypothesis testing and confidence intervals. A generic CC notation *r* will be used until a specific one is required.

Without loss of generality, we let $m = 0$ and then $Y = s\, Z$ with $E(Y) = 0,\ Var(Z) = 1$, and, as before, the vector $k = E(Z^o)$ has entries which are the expectations of the standardized order statistics. Assume it is desired to test $H_0 : s = s_0$ versus $H_a : s > s_0$. If H_0 is true, the random variable $r(k, (Y^o - s_o k)) = r(k, (\frac{Y^o}{s_o} - k)) =$ $r(k, (Z^o - k))$ will have a null distribution. If s_0 is too small (i.e., Ha is true), a plot of k and the order statistics from a random sample divided by the hypothesized standard deviation, y^o / s_0, will produce a line that is too steep; or, equivalently, the vector $(y^o / s_0) - k$ will not be centered at zero but, in general, will have more positive values. In any case, $r(k, (\frac{y^o}{s_0} - k))$ will tend to be large. Equivalently, if $z^o = y^o / s_0$ and $r(k, z^o - s\,k) = 0$ is solved for s with solution $s(z^o)$, then $s(z^o)$ will also tend to be larger than one. Thus large positive values will lead to rejection. If s_c is the slope with H_0 true such that $r(k, z^o - s_c k) = r_{a/2}$ where $r_{a/2}$ is the upper $a/2$ point for correlation coefficient *r*, then s_c is the upper $a/2$ critical value for the statistic *s*. Because of the monotonic power function property shown below it is true that

$$r(k, z^o - k) > r_{a/2} \Leftrightarrow s(z^o) > s_c .$$

-

For testing H_0 against $H_a : s < s_0$, if H_a is true, the vectors *k* and $z^o - k$ will tend to produce too negative a correlation value, and the rejection region will be for negative values. Again this rejection region has its counterpart in the $s(z^o)$ statistic.

Because of the monotonicity of $r(k, y^o - s\,k)$ as a function of *s*, the hypothesis test will have the necessary monotonic power function property. Reconsider the case $H_0 : s = s_0$ versus $H_a : s > s_0$. Then if $d \geq 1$, for the test to have monotonic power it is required that $r(k, (dy^o) - k) \geq r(k, y^o - k)$. However, $r(k, (dy^o) - k) = r(k, dy^o - k) = r(k, y^o - \frac{1}{d}k) \geq r(k, y^o - k)$ since $0 < 1/d \leq 1$ and *r* as a function of the coefficient of the vector *k* is decreasing (Gideon, 1992).

To form confidence intervals, let $u_{a/2} > 0$ and $l_{a/2} < 0$ be the upper and lower critical points of the null distribution (assuming in the case of a NPCC, $a/2$ is chosen to be one of the "natural levels" of the null distribution (Randles and Wolfe, 1979, p. 122)). Then determine s_l, the lower endpoint, and s_u, the upper endpoint of the 1 – α confidence interval by solving:

$$r(k, \frac{y^o}{s_l} - k) = u_{a/2}, \text{ and}$$

$$r(k, \frac{y^o}{s_u} - k) = l_{a/2}.$$

-

Note that $s_l < s(y^o) < s_u$ where $s(y^o)$ is the estimate of s depending on the particular r that has been chosen.

(Gee, 2002) investigated the use of five different correlation coefficients to determine confidence intervals for s. These included two parametric methods: Pearson's r_p and absolute value (see web site Publication 1) and three nonparametric methods: Kendall 's t, *GDCC*, and the modified footrule, or Gini's method, (David, 1968). Upper and lower critical points, $r_{a/2}$ and $-r_{a/2}$, of each null distribution were developed via computer simulations for a = 0.1 and a = 0.05 for samples of size 5 through 100. Tables were constructed giving the 0.025, 0.05, 0.95, and 0.975 quantile values.

(Gee, 2002) obtained, by simulations, the null distributions of $r(k, Z^o - k)$ for the five aforementioned CCs for sample sizes 5 through 100. He included histograms of selected null distributions with sample sizes of 5, 10, 30, 50, 75, and 100. For the two parametric CCs, absolute value and Pearson's r_p, with *n* = 5, the histograms are roughly ∪-shaped with a positive bias. As the sample sizes increase, the histograms are left-skewed. For the absolute value CC, the middle 50% of the data for sample size *n* = 30 falls in $(-0.212, 0.457)$ and for *n* = 100 in $(-0.239, 0.402)$, while for Pearson's r_p the middle 50% for *n* = 30 falls in $\left(-0.192, 0.618\right)$ and for *n* = 100 in $(-0.212, 0.547)$. Since the NPCCs take on a finite number of values, larger sample sizes are required to more readily see patterns. For sample sizes of 30 and above, both Gini's and Kendall's t are slightly skewed to the left while *GDCC* is more symmetric. For Gini's, the middle 50% of the data for sample size *n* = 30 falls in $(-0.249, 0.458)$ and for *n* = 100 in $(-0.280, 0.410)$; for

Kendall’s t, the middle 50% for n = 30 falls in $(-0.255, 0.370)$ and for n = 100 in $(-0.262, 0.334)$; while for *GD*CC the middle 50% for n = 30 falls in $(-0.267, 0.333)$ and for n = 100 in $(-0.260, 0.320)$. The minimum number of simulations used was 100,000.

8. A Numerical Example

An example from (Nemenyi, Dixon, White, and Hedstrom, 1977, p. 240) and (Iglewicz, 1983, pp. 408-410) is used in order to compare the performance of *GDCC* and Kendall's t to the robust estimators of scale that appear in these books. It is readily apparent that these two NPCCs used as scale estimators are among the best of the robust estimators. Two samples of SAT scores are used: one sample from a rural population with one outlier and a second sample from an urban population. The primary interest is in the comparison of the dispersions between the samples. (Iglewicz, 1983, p. 410) shows that the ratio of the lengths of the boxplots of the urban SAT scores to the rural SAT scores is 2.01. Let s and s' be the classical least squares estimates of standard deviation for the rural SAT scores with and without the outlier. Table 5 contains various estimates of scale for the rural and urban SAT scores. For the rural scores the sample standard deviation changes from s = 120.37 to s' = 82.20. The NPCCs are s_{gd} = 104.76 and without the outlier 87.24; for Kendall's t, s_t = 110.04 and changes to 94.70. Both have much smaller changes than the classical estimates of standard deviation. As is seen from Table 5, the ratios of the scales of urban to rural for *GDCC* is 2.06 and for the Kendall’s t, it is 1.82. Thus, the robustness of the NPCCs leads to reasonable results without worrying about the outlier. This robustness feature is one of the main reasons for using NPCCs as location and scale estimators. The other entries in the table are taken from Iglewicz with *AD*, the mean

-

absolute deviation; *MAD*, the median absolute deviation (both deviations from the median); d_F, the difference between upper and lower quartiles; s_{bi}, the *M* biweight estimator of scale using the biweight estimate of location.

Table 5: Comparisons of different scale estimates for the two samples of SAT scores

Estimator	Rural Students(1)	Urban Students(2)	Ratio (2)/(1)
s	120.37	176.58	1.47
s'	82.20	176.58	2.15
AD	81.62	144.54	1.77
MAD	47.00	149.00	3.17
d_F	85.00	277.00	3.26
s_{bi}	98.14	178.99	1.82
s_{gd}	104.76	215.48	2.06
s_t	110.04	200.06	1.82

Entries for s, s', *AD*, *MAD*, d_F, s_{bi} are from (Iglewicz, 1983, pp. 410, 424)

9. Summary and Comments

This paper is the third in a series of papers promoting the use of the geometry induced by a CC as a general estimating tool. In implementing these procedures, Pearson's r_p, for the most part, parallels least squares procedures. For NPCCs *GDCC* and Kendall’s t, a computer component is needed with the maximum-minimum tie breaking method first

-

suggested in (Gideon and Hollister, 1987). Computer programs can be written fairly easily for Kendall's t since a closed form regression estimation formula exists. For *GDCC* a *C*-language program has been written which combines the tie breaking procedure for the CC calculation and simple linear regression so that together these can be used in a variety of situations; *e.g*, multiple linear regression. It is a remarkably fast routine that makes higher-level problems feasible. An applied user would need a statistical software package to implement these ideas for general use. For this to happen, it needs to be ascertained how the "system of estimation" provided by a particular CC compares, say, to least squares. These authors are convinced that since *GDCC* is an "area equalizer" type estimator, it has the properties needed in real data analysis. However, the research effort needed to compare systems is beyond the means of the authors; the authors are thankful for *S-Plus* that makes available efficient research languages that have allowed for the progress thus far. Master's students and a few Ph.D. students have provided inspiration and technical help.

Censored data problems have been minimally studied and the ideas of this paper extend to such problems. They have not been included because of length considerations and hopefully will be included in a later paper. The following code is used to find the max and min upon which an NPCC is computed and then averaged to obtain a unique CC value. The paired vector data is (x, y). Other computer languages can implement the following commands:

rank(x) gives the ranks of *x*, usual average ranks used for ties;

order(y, x) contains indices of data elements (y) in ascending order (first integer is subscript of smallest element of y) with *y*-ties broken by values in x;

n is length(x), the size of the vectors;

<- denotes: evaluate right-hand side and put in left-hand side;

n1n <- 1:n puts the integers 1,2,..,n in n1n;

nn1 <- n:1 puts the 1st n integers in reverse order in nnl;

x[order(y)] gives the ordering of x that corresponds to the ascending ordered y.

	max	min
1	xt <- x[order(y,x)]	xr <-n+1-rank(x)
2	uv <- n1n[order(xt,n1n)]	xt <- x[order(y,xr)]
3	----------------------	uv1 <- n1n[order(xt,nn1)]

The vector *uv* is now a permutation of the first *n* positive integers upon which a NPCC can be computed. The vector *uv* corresponds to the ranks of the *y*-data after the *x*-data has been ordered. Compute the NPCC on *uv* for the max and *uv1* for the min and average the two results; this computation is one of the standard algorithms to compute Kendall's t.

As an example, let *x* = (1, 5, 6, 6, 3, 6, 1, 5, 4, 5, 6, 3, 3) and *y* =(7, 2, 6, 5, 6, 6, 2, 7, 6, 2, 6, 1, 4) then *uv* = (2, 12, 1, 5, 7 ,8 ,3, 4, 13, 6, 9, 10, 11) and *uv1* = (13, 4, 11, 5, 1, 10, 12, 3, 2, 9, 8, 7, 6). Kendall's t on *uv* is 0.4102564 and on *uv1* is -0.1794872 and the value of t for the *x*-*y* data is the average of these two values which is 0.115385. One can check the logic by hand on the *x*-*y* data by sorting and breaking tied ranks to either maximize or minimize the final result. *GDCC* = 1/3 for *uv* and 0 for *uv1* so the average is 1/6. For Pearson's, $r_p = 0.2089$.

-

There is one last observation for Kendall's t. Let the usual location two-sample problem be set up through regression; $i.e.$, 0 and 1 are the *x*-values and the *y*-values are the two sets of data plotted in the vertical directions. Then the slope of the scale regression line (1) with Kendall's t is the usual Hodges-Lehmann nonparametric location estimate, $\underset{i,j}{median}\{x_i - y_j\}$. This may also be true, in general, for *GDCC*, but, at this time, what is known is that it was always true for all the examples examined.

-

References

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Boston: Wadsworth-Duxbury Press.

D’Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. Biometrika 58:341-348.

-----(1973). Monte Carlo power comparison of the W and D tests of normality. Communications in Statistics 1:545-551.

David, H. A. (1968). Gini's mean difference rediscovered. Biometrika 55:573-575.

-----(1970). *Order Statistics*. New York: John Wiley & Sons.

Downton, F. (1966). Linear estimates with polynomial coefficients. Biometrika 53:129-141.

Gee, J. (2002). Using correlation coefficients and order statistics to estimate sigma. unpublished Master’s thesis, University of Montana, Dept. of Mathematical Sciences.

Gibbons, J. D. and Chakraborti, S. (1992). *Nonparametric Statistical Inference* (3rd ed.). New York: Marcel Dekker, Inc.

Gideon, R. A. (1992). The correlation coefficients. unpublished paper (URL: http://www.math.umt.edu/gideon/corrcoef.pdf), University of Montana, Dept. of Mathematical Sciences.

-

Gideon, R. A., and Hollister, R. A. (1987). A rank correlation coefficient resistant to outliers. Journal of the American Statistical Association 82: 656-666.

Gideon, R. A., and Rummel, S. E. (1992). Correlation in simple linear regression," unpublished paper (URL: http://www.math.umt.edu/gideon/CORR-N-SPACE-REG.pdf), University of Montana, Dept. of Mathematical Sciences.

Harter, H. L. and Balakrishnan, N. (1996). *CRC Handbook of Tables for the Use of Order Statistics in Estimation*. New York:CRC Press.

Hettmansperger, T. (1984). *Statistical Inference Based on Ranks*. New York: John Wiley & Sons.

Iglewicz, B. (1983). Robust scale estimators and confidence intervals for location. In: Hoaglin, D. C. Mosteller, F. Tukey, J. W. ed., *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley & Sons.

Looney, S. W. and Gulledge, T. R. (1985). Use of the correlation coefficient with normal probability plots. The American Statistician 39:75-79.

Nemenyi, P., Dixon, S. K., White, N. B., and Hedstrom, M. L. (1977). *Statistics from Scratch*. San Francisco: Holden Day, Inc.

Randles, R. H. and Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*. New York: John Wiley & Sons.

-

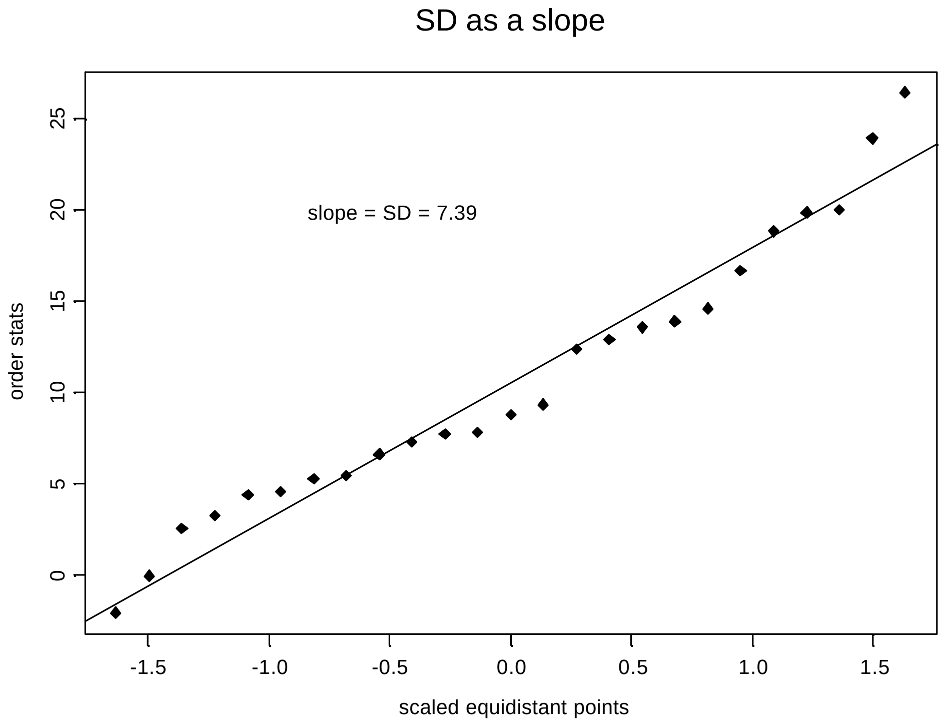
Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Venables, W.N. and Ripley, B.D, (1994). *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.

Acknowledgments

R. A. Gideon is indebted to many people for support over the years; they are listed on the web site. In particular, he thanks the following Masters' students for their interest and help on ideas developed in this paper: Bruder, J., Lee, L. C., Li, H., and Thiel, M. Their curiosity and enthusiasm have kept this research alive. Also Carol Ulsafer gave valuable editorial assistance.

-


SD as a slope
slope = SD = 7.39
order stats
0
5
10
15
20
25
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
scaled equidistant points

-