2-2011

Location and Scale Estimation with Correlation Coefficients

Rudy Gideon
University of Montana, Missoula

Adele Marie Rothan

Communications in Statistics—Theory and Methods, 40: 1561–1572, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610921003694430

Location and Scale Estimation with Correlation Coefficients

RUDY GIDEON[1] AND ADELE MARIE ROTHAN[2]

[1]Emeritus, Department of Mathematical Sciences, University of Montana, Missoula, Montana, USA
[2]Department of Mathematical Sciences and Physics, St. Catherine University, St. Paul, Minnesota, USA

This article shows how to use any correlation coefficient to produce an estimate of location and scale. It is part of a broader system, called a correlation estimation system (CES), that uses correlation coefficients as the starting point for estimations. The method is illustrated using the well-known normal distribution. This article shows that any correlation coefficient can be used to fit a simple linear regression line to bivariate data 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. 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 discussed for the scale parameter; both normal and Cauchy distributions are covered.

Keywords Confidence intervals; Hypothesis testing; Robust estimates; Simple linear regression.

Mathematics Subject Classification Primary 62G05, 62G08; Secondary 62G10, 62G30, 62G35.

1. Introduction

This article uses three correlation coefficients (CC): Pearson's r_p, Kendall's τr_k, and Greatest Deviation Correlation Coefficient (GDCC or r_{gd}), as defined in 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; and Kendall's τ, a well-known nonparametric correlation coefficient (NPCC). 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

Received October 1, 2006; Accepted December 23, 2009

Address correspondence to Rudy Gideon, Emeritus, Department of Mathematical Sciences, University of Montana, Missoula, MT 59812, USA; E-mail: GideonR@mso.umt.edu

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 $Y = \mu + \sigma Z$, where Z is normal with mean 0 and standard deviation 1, so $Y \sim N(\mu, \sigma)$. Then for the order statistics $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$, $Y_{(i)} = \mu + \sigma Z_{(i)}$ and $E(Y_{(i)}) = \mu + \sigma E(Z_{(i)})$. Let $k_i = E(Z_{(i)})$, $i = 1, 2, \ldots, n$. From the symmetry of the standard normal, note that $\sum k_i = 0$. Randles and Wolfe (1979) next defined $D(\mu, \sigma) = \sum_{i=1}^{n} (Y_{(i)} - (\mu + \sigma k_i))^2$ The estimators $\hat{\mu}$ and $\hat{\sigma}$ that are found to minimize D are unbiased for μ and σ, respectively.

This solution is next related to Pearson's r_p. Again, 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 sample order statistics $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$. The measure of variability is now defined via simple linear regression, but on ordered data. The slope of the regression is a measure of the variability and in particular estimates standard deviation directly. If the following equation is solved for s using any r, then s estimates σ:

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

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 μ; in fact, these latter two estimates are identical to the ones coming from $D(\mu, \sigma)$. From Publication 2, *Correlation in Simple Linear Regression*, on the website (www.umt.edu/math/People/Gideon.html), with $x = k$ and $y = y^o$, the regression Eq. (1) of that article becomes the above (1), called the scale form of the regression equation. The solution is $s = \frac{\sum k_i y_{(i)}}{\sum k_i^2}$ and mean(*res*) $= \bar{y} - s\frac{\sum k_i}{n} = \bar{y}$. Note that the usual estimate of the mean is obtained. The estimate of σ for the random variable S is unbiased because

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

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

$$k = \frac{6}{(n+1)\sqrt{\pi}} \left\{ \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} - \frac{n+1}{2} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\}, \tag{2}$$

The solution for s in Eq. (1) with this k is related to both Gini's mean difference (Randles and Wolfe, 1979; Hettmansperger, 1984) and a method of Downton (1966).

Gini's mean difference estimate of scale (David, 1968) is

$$G(y) = \frac{1}{\binom{n}{2}} \sum_{i<j} |y_{(i)} - (y_{(j)})|,$$

and Downton's estimate of scale for the normal distribution is

$$s_{dt} = \frac{\sqrt{\pi}}{\binom{n}{2}} \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) (y_i).$$

It can be shown that $s_{dt} = \frac{\sqrt{\pi}}{2} D(y)$ so that Gini and Downton are essentially the same, and both can be obtained from Eq. (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, i.e., using Eq. (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 σ to perform a test of normality. The estimate of σ from Eq. (1) with $k_i = E(Z_{(i)})$, $i = 1, 2, \ldots, n$ could also be used in the D'Agostino normal test of fit with this s replacing the classical estimate. Another test of the normality assumption using Pearson's r_p is given in Looney and Gulledge (1985).

An interpretation of the usual $SD = \sqrt{\frac{\sum(y_i - \bar{y})^2}{(n-1)}}$ as the slope of a straight line is next used as a transition to a more geometrical view of scale estimates. Again consider the ordered data $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ and let constant $c = \sqrt{12/(n(n+1))}$. The choice of this c becomes apparent in the development. Think of the horizontal axis points as the vector h transpose, that is, $h' = \left(-\frac{n-1}{2}, -\frac{n-3}{2}, -\frac{n-5}{2} \cdots 0,\ 1, \cdots \frac{n-5}{2}, \frac{n-3}{2}, \frac{n-1}{2}\right)$. 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. The multiplication by c is merely a change of scale to keep the range of the plot on the horizontal axis, regardless of the data set, roughly between $\pm\sqrt{3}$ for all n, while keeping points equidistant.

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 horizontal line at $\bar{y}$ measures its departure from that line, while the *SD* is an overall measure of departure from the horizontal.

With this motivation, a straight line with slope b and intercept $\bar{y}$ is now constructed through the points (ch, y^o). To achieve the goal, the rescaled vector bch must have components whose cumulative squared distance from the $\bar{y}$ line are the same. In other words, b is chosen so that $\sum (y_i - \bar{y})^2 = \sum_{i=1}^{n} (bch_i)^2$. The line with slope b defines an angle θ with the horizontal having $\tan\theta = b$. Because $\sum_{i=1}^{n} h_i^2 = 2\sum_{j=1}^{\frac{n-1}{2}} j^2 = \frac{n(n-1)(n+1)}{12}$ and using the definition of c, the b that satisfies the above equation is $b = SD$. Figure 1 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 vector of vertical values is $10.56 + 7.39 * (ch)$. The slope b, which is the *SD*, represents the variation in the data. A steeper slope (or a larger θ) implies more variation and a 0 slope (or $\theta = 0$) indicates no variation.

The points (ch, y^o) can be used to illustrate the CES way of estimating σ. The scale regression Eq. (1) is solved for s using $k = ch$. Because Pearson's r_p is still being used the solution is labeled s_p. The equation is $r_p(ch,\ y^o - s_p ch) = 0$. It is straightforward to obtain $s_p = \sqrt{\frac{12}{n(n+1)}} \frac{1}{n-1} \sum_{i=1}^{n} h_i y_{(i)}$. For the data used in the figure, $s_p = 7.24$. Now solving Eq. (1) with r_{gd} gives the scale estimate 6.80. After a few computer runs it was clear that the estimations coming from 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

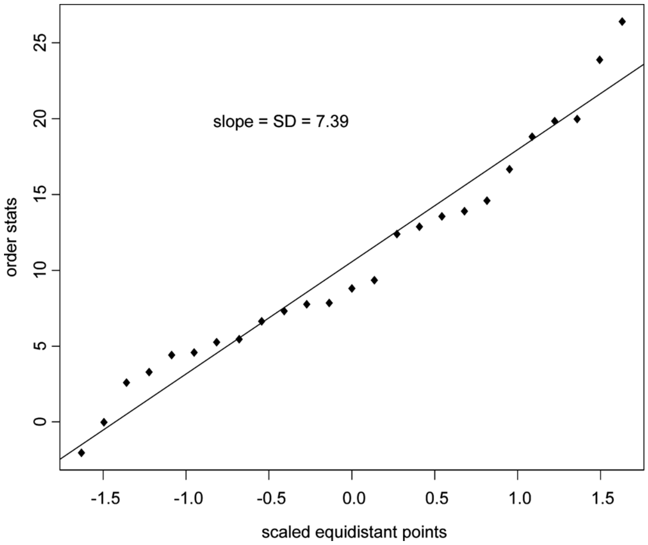

Figure 1. SD as a slope.

estimate does. His constant is $\frac{6}{(n+1)\sqrt{\pi}}$ as compared to c. The ratio of Downton's constant to c is $\sqrt{3(n+1)/(\pi n)}$, which is about 0.9772, assuming $(n+1)/n \cong 1$. 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 correlation coefficient can be used to estimate location and scale as in the example above. Equation (1) can be solved with any correlation coefficient, r, each giving an estimate of σ. 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 data.

2. Interpreting Equation (1)

When r_{gd} is used for r in Eq. (1), the solution s must be found numerically as no closed form solution is known, but for Kendall's τ, the equation $r_k(k, y^o - sk) = 0$ is satisfied by $s = median\left(\frac{y_{(j)}-y_{(i)}}{k_j-k_i}\right)$ (see website Publication 2 for Tau and R program for GDCC). Because both of these NPCCs are discrete, 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 $r = \tau$ or for any nonparametric r.

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

- for Pearson's r_p, $s(y)$ is a continuous function and (1) has a closed form solution;
- for GDCC, $s(y)$ is a nonincreasing step function based on a NPCC and (1) has only a numerical solution;
- for Kendall's τ, $s(y)$ is a nonincreasing 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, $s(Y + d * 1) = s(Y)$, where d 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) used the term "equivariant" for statistics that transform properly. Note that $s(Y)$ is scale equivariant, i.e., if $d > 0$ is a constant and $X = dY$ is a scale change, then $s(dY) = ds(Y)$, as is now easily shown. Because $r(k, X - s(X)k) = 0$ for any vector k of equidistant points and CCs are scale invariant, seeing that $r(k, X - ds(Y)k) = r(k, dY - ds(Y)k) = r(k, Y - s(Y)k) = 0$, verifies that $ds(Y)$ is $s(X)$, that is, $s(dY) = ds(Y)$. The evenness argument is set out in a proposition.

Proposition 3.1. *Given data from a symmetric distribution about* 0 *and a solution s of Eq.* (1) *using any NPCC r,* $s(Y) = s(-Y)$.

Proof. Since the distribution is symmetric about 0, $k_{n+1-i} = -k_i$, $i = 1, 2, \ldots, n$ and for the vector k, $(-k)^o = k^o = k$. It is also true that

$$(-Y)^o = \begin{pmatrix} -y_{(1)} \\ -y_{(2)} \\ \vdots \\ -y_{(n-1)} \\ -y_{(n)} \end{pmatrix}^o = \begin{pmatrix} -y_{(n)} \\ -y_{(n-1)} \\ \vdots \\ -y_{(2)} \\ -y_{(1)} \end{pmatrix}.$$

Substituting $-y$ for y in Eq. (1) gives

$$0 = r(k, (-y)^0 - s(-y) * k) = r((-k)^0, \quad (-y)^0 - s(-y) * (-k)^0),$$

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 Eq. (1),

$$\begin{aligned} 0 &= r\left((-k), (-y) - s(-y) * (-k)\right) \\ &= r\left(-k, -\left(y - s\left(-y\right) * k\right)\right) = r\left(k, y - s\left(-y\right) * k\right); \end{aligned}$$

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

The scale estimate is obtained with the same k even if the data are not centered at 0 but are symmetric.

4. Motivation and Standard Properties of the CES Location Estimator of Y

Because CCs are the estimation tools in CES, 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 $1 \leq i \leq m$ and the y-data as (1, y_i) for $1 \leq i \leq n$. 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 let the $m+n$ dimension vector v be $(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n)'$. Treat c as the regressor variable and v as the response variable. Then the correlation coefficient 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 $\bar{y}-\bar{x}$ or $\bar{x}+slope=\bar{y}$. 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 $r_k(c, v-lc)=0$ 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_i are zero, so $l=median(y_j)$. As discussed in Gideon and Rummel (1992), if the x-data are all zero 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 $s=s_\tau$ or $s=s_{gd}$, the solution of $r(k, y-sk)=0$ when r is τ or GDCC, respectively.

Proposition 4.1. *For NPCCs τ and GDCC and for a symmetric distribution, $l(y)=median(y^o-sk)$ is an odd translation statistic.*

Proof.

$$\begin{aligned} l(-y) &= median\left((-y)^o - sk\right) = median\left((-y)^o - s\left(-k\right)^o\right) \\ &= median\left((-y) - s\left(-k\right)\right) = -median\left(y - sk\right) = -l(y) \end{aligned}$$

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 Eq. (1) using Kendall's τ, it is possible to make a closer examination of its scale and location estimates.

Let the elementary slopes be $l_{ji}=\frac{Y_{(j)}-Y_{(i)}}{k_j-k_i}$, for $1 \leq i < j \leq n$ where $k_i=E(Z_{(i)})$. Now $E(l_{ji})=\frac{E(Y_{(j)})-E(Y_{(i)})}{k_j-k_i}=\frac{(\mu+\sigma k_j)-(\mu+\sigma k_i)}{k_j-k_i}=\sigma$. Each l_{ji} can be considered a random observation from a population with mean σ, therefore, $E(mean(l_{ji})) = \sigma$. However, to be unbiased, the scale estimator, $s_\tau(Y) = median(l_{ji})$, depends on the symmetry of the distribution of the correlated l_{ji}. The quantity s_τ 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, σ. Simulations show that s_τ appears to have a slight positive bias in estimating *SD*. If $res_i = Y_{(i)} - s_\tau k_i$, $i = 1, 2, \ldots, n$ and $E(s_\tau(Y)) = \sigma^+ > \sigma$, then $E(res_i) = E(Y_{(i)} - s_\tau k_i) = (\mu + \sigma k_i) - \sigma^+ k_i = \mu + (\sigma - \sigma^+)k_i$. Because each residual, res_i, has expectation possibly slightly less than μ for $k_i > 0$, but slightly greater for $k_i < 0$, the expectation of the median of the residuals is approximately μ. 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, Sec. 1.3) can be used to show that $s_\tau(y)$ and $l_\tau(y)$ are uncorrelated statistics. Of course, for the normal distribution, the classical estimate of σ 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, $l_\tau(y)$, is symmetrically unbiased. In the CES, it is necessary to first estimate the scale and then the location.

Assume $Y^* - \mu \overset{d}{=} \mu - Y^*$, i.e., Y^* is symmetric about μ; the usual "equal in distribution" notation has been used. Then without loss of generality, $Y = Y^* - \mu$ is symmetric about zero. The distribution function $F(y)$ is

$$F(y) = P(Y \leq y) = P\left(Z \leq \frac{y - \mu}{\sigma}\right).$$

Because $\mu = 0$, $Y_{(i)} = \sigma Z_{(i)}$, $i = 1, 2, \ldots, n$. The estimate of the standard deviation with Kendall's τ, s_τ, is

$$s_\tau = \underset{i<j}{median}\left(\frac{y_{(j)} - y_{(i)}}{k_j - k_i}\right) \quad \text{where } k_i = E(Z_{(i)}).$$

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

Earlier it was shown in Proposition 3.1 that $s(Y) = s(-Y)$, but it is constructive to show this again specifically for Kendall's τ; that is, that $s_\tau(y) = s_\tau(-y)$. Let $X = -Y$ or for a random sample $x_i = -y_i$. Then for order statistics, $x_{(i)} = -y_{(n+1-i)}$, $i = 1, 2, \ldots, n$ and

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

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

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

Proposition 4.2. *Kendall's τ estimate of the median of a symmetric distribution has a symmetric distribution about the true population median; that is, $l_\tau(-y) = -l_\tau(y)$. In this case the mean and median are equal.*

Proof. The estimate of the population median based on the residuals of the scale estimate is $l_\tau(y) = median(y_{(j)} - s_\tau(y)k_j)$. As above, let $X = -Y$. Then:

$$\begin{aligned} l_\tau(-y) = l_\tau(x) &= median\left(x_{(j)} - s_\tau(x)\,k_j\right) \\ &= median\left(-y_{(n+1-j)} - s_\tau(x)\left(-k_{n+1-j}\right)\right) \\ &= -median\left(y_{(n+1-j)} - s_\tau(x)\left(k_{n+1-j}\right)\right) \\ &= -median\left(y_{(n+1-j)} - s_\tau(-y)\left(k_{n+1-j}\right)\right) \\ &= -median\left(y_{(n+1-j)} - s_\tau(y)\left(k_{n+1-j}\right)\right) \quad \text{because } s_\tau(y) = s_\tau(-y) \\ &= -median\left(y_{(j)} - s_\tau(y)\left(k_j\right)\right) \\ &= -l_\tau(y). \end{aligned}$$

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

5. Scale and Location Estimates using GDCC and Kendall's Tau

A comparison was made of estimates of the location and scale from simulations of the normal distribution with and without outliers. The classical mean, median, and standard deviation were compared to the comparable estimates via GDCC and Tau. Bias was assessed by comparing the means of the estimates for both location and scale parameters. Mean square error was computed to measure the variation of the estimates.

For the case of no outliers the classical standard deviation is very slightly better than the GDCC and Tau methods. Tau and GDCC scale estimates were biased slightly upwards by about 4%. However, when a few outliers were randomly added to the data, both GDCC and Tau were far better than the classical standard deviation. This was true both for average values and in the variation of the estimates. For good data, the variation of the standard deviation estimate was about 30% higher for GDCC and 15% higher for Tau than the classical standard deviation. However, with a few outliers, the classical standard deviation was three times more variable than the nonparametric correlation coefficient approach. GDCC had the least variable SD estimate with the least bias.

Recall that to estimate location for both GDCC and Tau, the median is computed using the residuals after the estimate of the SD. Surprisingly, this new method was superior to the standard median method. It appears to be unbiased and had a variation comparable to the variation in the mean rather than the larger variation of the usual median. With outliers randomly added, the CES median method and the standard median were far more robust, of course, than the mean. The variation of the median and CES median methods were about the same. Details are omitted here due to space constraints. The interested reader may consult Gideon and Hollister (1987) for background material on GDCC and website Publication 5 for information on the use of Tau.

6. Hypothesis Testing and Confidence Intervals for σ

There is an acute need for a better scale analysis because most scale tests under the normality assumption lead to unreliable results. 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.

Without loss of generality, we let $\mu = 0$ and then $Y = \sigma Z$ 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 $H_0 : \sigma = \sigma_o$ vs. $H_a : \sigma > \sigma_o$. If H_0 is true, the random variable $r(k, (Y^o - \sigma_o k)) = r\big(k, \big(\frac{Y^o}{\sigma_o} - k\big)\big) = r(k, (Z^o - k))$ will have a null distribution. If σ_o is too small (i.e., H_a is true), a plot of k and the order statistics from a random sample divided by the hypothesized standard deviation, y^o/σ_o, will produce a line that is too steep; or, equivalently, the vector $(Y^o/\sigma_o) - k$ will not be centered at zero but, in general, will have more positive values. In any case, $r\big(k, \big(\frac{Y^o}{\sigma_o} - k\big)\big)$ will tend to be closer to positive one. Equivalently, if $z^o = Y^o/\sigma_o$ 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 large. Thus, large positive values will lead to rejection of the null hypothesis.

To perform tests of significance and construct confidence intervals, first draw random samples (1,000 or so) from the standardized distribution being considered (normal and Cauchy are used here) and then tabulate the distribution of σ, the solutions of $r(k, z^o - \sigma k) = 0$ to closely estimate the null distribution. This can be done for many rs; in these examples r_{gd} was used with a sample size of $n = 25$. For a desired α, determine $w_{\alpha/2}$ and $w_{1-\alpha/2}$ from the distribution constructed from the simulations. For $\alpha = 0.05$, the following was obtained:

- for N(0, 1): $w_{.025} = 0.672$ and $w_{.975} = 1.401$;
- for Cauchy with median 0 and scale factor 1: $w_{.025} = 0.604$ and $w_{.975} = 2.066$.

The construction of the confidence interval is explained first. Once a confidence interval has been obtained the usual analogy with hypothesis testing can then be used for a hypothesis test.

The point estimate of σ, the scale factor, is $\hat{\sigma}$ where $r_{gd}(k, y^o - \hat{\sigma}k) = 0$. Now, let σ_l and σ_u denote the lower and upper points of the confidence interval, respectively. These are found by solving the equations $r_{gd}(k, y^o - \sigma_l w_{1-\alpha/2} k) = 0$ and $r_{gd}(k, y^o - \sigma_u w_{\alpha/2} k) = 0$. It might help to write, say the first equation, as $r_{gd}(k, y^o/\sigma_l - w_{1-\alpha/2} k) = 0$. In other words, the data is being standardized to correspond to the upper point $w_{1-\alpha/2}$ of the standardized distribution. This approach is analogous to the classical chi-square methods to obtain the confidence interval but now uses an implicit function. Note that the Cauchy is viable when a rank based correlation coefficient is used, but not Pearson's r_p where moments must exist. The interval (σ_l, σ_u) is a $1 - \alpha/2$ confidence interval for scale factor σ. The k used was an approximation for the expected values of the order statistics (Gibbons and Chakraborti, 1992). Specifically, let F be the cumulative distribution function of the selected standardized (0, 1) random variable and let $v = \big(\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}\big)$. $F^{-1}(v)$, an ordered n dimensional vector, was used as k. This is easy to implement in R or Splus. These ideas were checked out for both the normal and Cauchy distributions with different αs. The confidence intervals performed exactly as required.

To test $H_o : \sigma = \sigma_o$ vs. some alternative, the rejection regions are one or both of $\sigma < \sigma_l$, $\sigma > \sigma_u$. The R instruction to run the above is $\hat{\sigma} = uniroot(GDslp, c(-20, 20), x = k, y = y^o)\$root$, where y^o is the sorted data, *GDslp* is an R program for the slope of a simple linear regression line using GDCC. Note that for σ_l, $x = (w_{1-\alpha/2}k)$ and for σ_u, $x = (w_{\alpha/2}k)$.

7. A Numerical Example

An example, including a modified table from Nemenyi et al. (1977, p. 240) and Iglewicz (1983, pp. 408–410), is used in order to compare the performance of GDCC and Kendall's τ 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) showed that the ratio of the lengths of the boxplots of the urban SAT scores to the rural SAT scores is 2.01 and the author indicates that this ratio is best. With the outlier the classical least squares estimates of standard deviation for the rural SAT scores is $s = 120.37$, and without the outlier it is $s' = 82.20$. Without the outlier deletion, s gives a poor estimate of the dispersion ratio. The NPCCs are $s_{gd} = 104.76$ and without the outlier 87.24; for Kendall's τ, $s_\tau = 110.04$ and without the outlier changes to 94.70. Both have much smaller changes than the classical estimates of standard deviation. As is seen from Table 1, the ratios of the scales of the original urban to rural data for GDCC is 2.06 and for Kendall's τ, it is 1.82. Note that GDCC gives a good result without examining the outliers. This robustness feature is one of the main reasons for using NPCCs as location and scale estimators. The other entries in Table 1 are taken from Iglewicz (1983).

The technique of Eq. (1) can be extended to the estimate of the ratios of any two standard deviations where the sample sizes are equal. For example, for the SAT data let x^o be the sorted SAT rural data and y^o be the sorted SAT urban data. Then the ratio σ_x/σ_y can be estimated directly by solving $r(x^o, y^o - sx^o) = 0$ for s. When this was done using r_{gd} the result was 1.98, close to the 2.06 in Table 1. This estimate requires no intermediate steps. The idea is also useful in multiple linear regression to directly estimate σ_{res}/σ_y which is a key term in the analysis.

Table 1
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_τ	110.04	200.06	1.82

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

8. Summary and Comments

This article is part of a series of articles (Gideon, 2007; Gideon and Hollister, 1987) promoting the use of CCs as a general estimating tool. In implementing these procedures, it was found that Pearson's r_p, using CES ideas, parallels least squares procedures. For NPCCs GDCC and Kendall's τ, a computer component is needed with the maximum-minimum tie breaking method first suggested in Gideon and Hollister (1987). A short and easy to use R program to compute GDCC and its slope for use in Eq. (1) is given on the website. Computer programs can be written fairly easily for Kendall's τ since a closed form regression estimation formula exists. 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 correlation coefficient compares, say, to least squares. The 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 Splus or R 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.

One example is given so that the max-min tie-breaking algorithm can be seen on data with ties. It is important because without it, ties would make the CES untenable.

As an example, let $x = (1, 5, 6, 6, 3, 6, 1, 5, 4, 5, 6, 3, 3)$ and $y = (7, 2, 6, 5, 6, 6, 2, 7, 6, 2, 6, 1, 4)$. To implement the max-min method, the x data is ordered and replaced by ranks and paired with y data in which first ranks, both for x and y, are chosen to maximize the correlation and then second ranks are chosen to minimize the correlation. This is done within the restrictions of the tied data. The x data is then the unique ranks 1 to n and for y the max algorithm gives (2, 12, 1, 5, 7, 8, 3, 4, 13, 6, 9, 10, 11) and the min algorithm gives (13, 4, 11, 5, 1, 10, 12, 3, 2, 9, 8, 7, 6). Kendall's τ on the first is 0.4102564 and on the latter is -0.1794872 and the value of τ 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 the max and 0 for the min so the average is $1/6$. For Pearson's, $r_p = 0.2089$.

There is one last observation for Kendall's τ. 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 from Eq. (1) with Kendall's τ is the usual Hodges-Lehmann nonparametric location estimate, $median_{i,j}\{x_i - y_j\}$. 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.

Acknowledgments

R. A. Gideon is indebted to many people for support over the years; please see http://www.umt.edu/math/People/Gideon.html. In particular, he thanks the following Masters' students for their interest and help on ideas developed in this article: J. Bruder, L. C. Lee, H. Li, and M. Thiel. Their curiosity and enthusiasm have kept this research alive. Also, Carol Ulsafer gave valuable editorial assistance.

Much of the work on CES is available on the website: www.umt.edu/math/People/Gideon.html. Some of the references will refer to particular articles posted at this website.

References

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

D'Agostino, R. B. (1973). Monte Carlo power comparison of the W and D tests of normality. *Commun. Statist.* 1:545–551.

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

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

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

Gideon, R. A. (2007). The correlation coefficients. *J. Mod. Appl. Statist. Meth.* 6:517–529.

Gideon, R. A., Hollister, R. A. (1987). A rank correlation coefficient resistant to outliers. *J. Amer. Statist. Assoc.* 82:656–666.

Gideon, R. A., Rummel, S. E. (1992). Correlation in simple linear regression, unpublished paper at http://www.umt.edu/math/People/Gideon.html, University of Montana, Department of Mathematical Sciences.

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., eds. *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley & Sons.

Looney, S. W., Gulledge, T. R. (1985). Use of the correlation coefficient with normal probability plots. *Amer. Statistician* 39:75–79.

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

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

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