

# On the Robustness of a Class of Naive Estimators

Howard Wainer  
Bureau of Social Science Research

David Thissen  
University of Kansas

A class of naive estimators of correlation was tested for robustness, accuracy, and efficiency against Pearson's  $r$ , Tukey's  $r_t$ , and Spearman's  $r_s$ . It was found that this class of estimators seems in some respects to be superior being less affected by

outliers, reasonably efficient, and frequently more easily calculated. The definition and details of the use of these naive estimators are the subject of this paper.

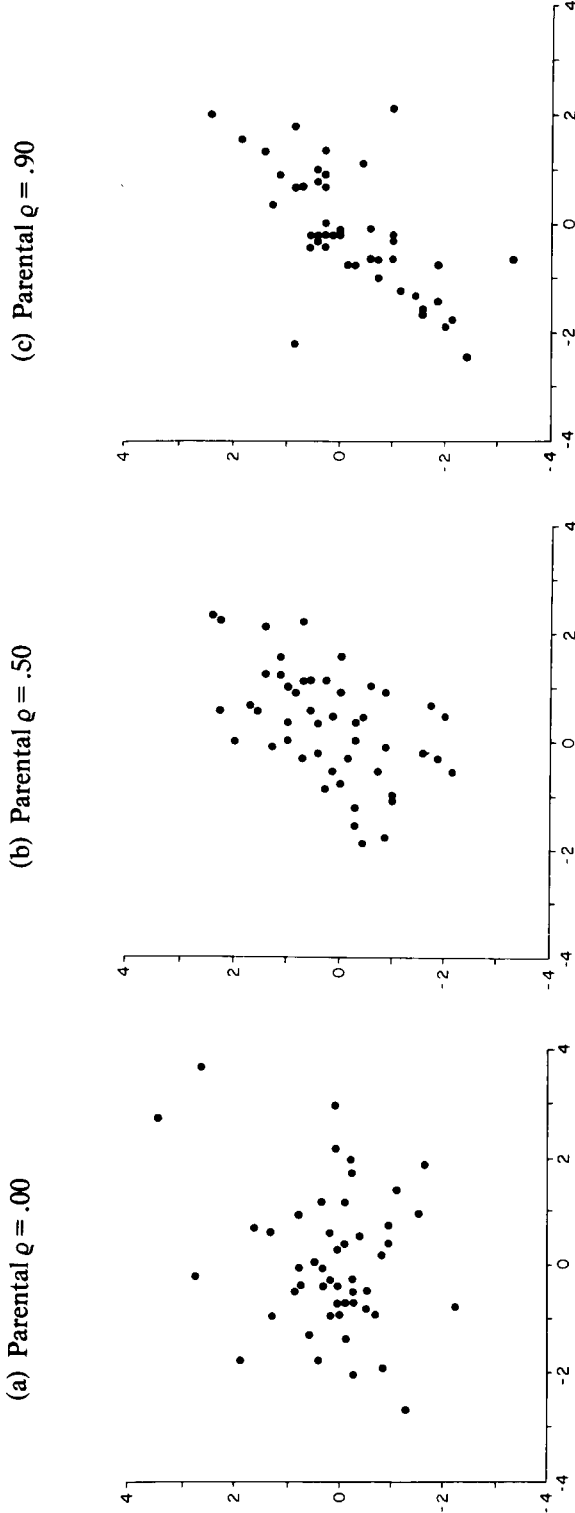
The distribution theory associated with the Pearson product-moment correlation ( $r$ ) is quite fragile with respect to violation of its underlying assumption of bivariate normality. In practice among social scientists, this fragility is almost universally ignored, although statisticians have proposed numerous schemes for protection (e.g., Hogg, 1974). Figure 1 shows bivariate distributions which all have  $r = .5$  and sample size equal to 50. In each case, the application of the usual normal surface theory establishes that the correlation is statistically significant beyond the  $p = .05$  level. More precisely, a normal theory 95% confidence interval around the estimated population value of  $\rho$  is

$$\text{Probability } (.26 \leq \rho \leq .68) \geq .95. \quad [1]$$

As is clear from viewing the bivariate plots, the correlation in Figure 1a is probably small or non-existent in the population. The correlation in Figure 1b is probably aptly described by the confidence interval in Equation 1, and the confidence interval for the correlation in Figure 1c is much larger. These conclusions can be made quickly through inspection, yet it is misleading to use the usual statistical techniques. The question is, Why? The confidence intervals are constructed through the use of a standard error which, if the data are bivariate Gaussian, is purely a function of  $n$ ; thus, the confidence intervals use the normality assumption. The standard error is indirectly estimated and is incorrect. A more realistic estimate of the standard error can be obtained through Jackknifing (Wainer & Thissen, 1975), and this same technique can be used to get a more conservative estimate of the confidence interval. But this technique also gives a poor estimate of the population value of  $\rho$ . The question to be answered is thus, How can a better estimate be obtained?

**Figure 1**

Bivariate Distributions with  $r = .50$  ( $n = 50$ )



Before attempting to answer this question, it is worthwhile to describe how the three distributions in Figure 1 were generated. Each is the mixture of two Gaussian distributions in various proportions. Figure 1a has 80% of its points drawn from a “parent” bivariate Gaussian distribution with mean 0, standard deviation 1, and correlation 0 and 20% of its points drawn from a “contaminating” bivariate Gaussian distribution with mean 0, standard deviation 2, and correlation .98. Figure 1b is completely drawn from one standard normal distribution with  $\rho = .5$ . And Figure 1c has  $\rho = .9$  for the standard normal parental distribution, with 20% of its points drawn from a contaminating normal distribution with mean 0, standard deviation 2, and correlation = 0.0.

The distinction between parental and contaminating portions of the artificial data distributions is meant to be an analog to circumstances frequently encountered in data of interest to social scientists, in which sampling and experimental control are frequently less than ideal, sometimes including in the data observations that represent some population distinct from the majority population studied. Both the study of patterns present in the majority of the data and the detection and study of individual contaminating observations, or outliers, should be goals of the data analyst.

The position taken in this paper is that to detect outliers, the majority pattern from which the outlying observations diverge must first be established. Then the divergent observations can be studied against the background of the majority of the data. Thus, it is assumed that the investigator is interested in obtaining the best possible estimate for the population correlation represented by the majority of the data and an honest estimate of the estimate’s variability.

### Three Robust Estimators of Correlation

Although the interest of the authors is in a statistic that estimates the correlation representing the parental portions of these sampling distributions, these are not the only situations for which it is desirable that this statistic work. These samples simply provide an analog for the general class of distributions which are fatter tailed than normal. The particular situation of Gaussian distributions contaminated by Gaussian noise of higher variance has been investigated by Box and Tiao (1968). The general case of fat-tailed distributions requires more general robust estimates of correlation: Three such estimators will be discussed.

#### Spearman Rank-Order Correlation

The first candidate for a robust correlation coefficient is the Spearman rank-order correlation, which will be denoted  $r_0$ . Transforming this to be an estimator of  $\rho$  (for ranked normal data) by the relation

$$r_0 = 2 \sin \frac{\pi}{6} r_1 \quad [2]$$

gives a statistic which is rather robust with respect to outliers, but suffers drawbacks in terms of efficiency when the data are nearly normal (Wainer & Thissen, 1976).

#### Tukey Correlation

The second candidate for a robust measure of correlation is the Tukey correlation  $r_t$ , which is not quite as robust as  $r_0$ , but more efficient in Gaussian and near-Gaussian situations. It is more robust

than  $r$  and is a good compromise. This statistic uses some order statistics in its computation and is calculated as follows:

To begin, an estimator of scale  $s^*(x)$ , a robust estimator of the standard deviation, is defined. The observation vector  $\underline{x}$  is first ordered, so that  $x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n$ . Next, the "gaps"  $g_i$  are calculated as

$$\underline{g}_i = \underline{x}_{i+1} - \underline{x}_i \quad \underline{i} = 1, 2, \dots, n-1. \quad [3]$$

Also defined are a set of approximately Gaussian weights

$$\underline{w}_i = \underline{i}(n-i), \quad \underline{i} = 1, 2, \dots, n-1. \quad [4]$$

Then, the robust estimator of scale is

$$\underline{s}^* = \frac{\sqrt{\pi}}{n(n-1)} \sum_{i=1}^{n-1} \underline{w}_i \underline{g}_i \quad [5]$$

This statistic has been shown (Downton, 1966) to be an unbiased and efficient estimator of the standard deviation when the data are Gaussian and a robust estimator (Wainer & Thissen, 1976) when they are not.

This estimator of scale is used to construct a robust estimator of correlation as follows:

Define

$$\underline{x}'_i = \underline{x}_i / \underline{s}^*(x) \quad [6]$$

$$\underline{y}'_i = \underline{y}_i / \underline{s}^*(y) \quad [7]$$

These are analogous to z-scores. Then, the robust correlation is given by

$$\underline{r}_t = 1/4 \{ [\underline{s}^*(\underline{x}'+\underline{y}')]^2 - [\underline{s}^*(\underline{x}'-\underline{y}')]^2 \}. \quad [8]$$

### Naive Estimators

The third robust measure of correlation,  $r_p$ , summarizes a class of naive estimators which will be described in detail in the next section. Now, these three measures will be compared and contrasted with  $r$  and with one another with respect to their accuracy and efficiency both when the data are Gaussian and when they are not.

### Comparison of the Four Correlations

To compare the four measures of correlation,  $r$ ,  $r_o$ ,  $r_r$ , and  $r_p$ , they were tested under a variety of conditions: three levels of contamination (0%, 10%, and 20%), three levels of correlation in the parental distribution (.0, .5, and .9), and two sample sizes (50 and 100). Hence, the present study was

a  $3 \times 3 \times 2$  design with four measures of correlation tested in each cell of the simulation. There were 38 random samples drawn per cell, and the dependent variables reported were robust estimates of the mean in each cell for each correlation measure as well as robust estimates of the efficiency of each estimator. The contaminating distribution was bivariate Gaussian with mean 0, standard deviation 2, and correlation 0. The parent distribution was bivariate Gaussian with mean 0, standard deviation 1, and correlation  $p$ , with  $p$  taking on the three values specified earlier.

The numbers in Tables 1 and 2 are robust estimates. These were calculated in the following manner: In Table 1 the entries were obtained by first transforming the estimated correlations to  $z$ -scores through the use of the ordinary  $r$ -to- $z$  transformation. Next, the highest and lowest 10% of the scores were trimmed and the mean of the middle 80% calculated. This 10% trimmed mean was then back transformed to the metric of correlations. The efficiencies were obtained in a similar manner. The estimated correlations were transformed to  $z$ -scores, and then the lowest and highest 10% of the data were moved in and set equal to the next highest (or lowest) data point. This process is called "Winsorizing" (Dixon, 1960). The mean of these Winsorized data was then calculated and each data point was then subtracted from this 10% Winsorized mean. The squared residual from the Winsorized mean were then themselves Winsorized and summed. This Winsorized sum of squares around a Winsorized mean has been shown to be the appropriate error term associated with a trimmed mean (Tukey & McLaughlin, 1963). The displayed efficiencies are the ratios of these Winsorized sum of squares to the smallest of them within any row of the design. Thus, the estimator which varies the least has an efficiency of unity and all others have somewhat smaller efficiencies.

Approximately the same result holds if less robust summary statistics are used (such as the uncensored mean and variance); but since this paper has, as its main point, the fragility of least squares estimators, it would have been inappropriate to use them as the key summary statistics.

## Results

Table 1 gives the 10% trimmed means obtained in the simulation. It is clear that when there was no contamination, all estimators did reasonably well. Of course,  $r$  must be superior to the rest, since it is the maximum likelihood estimator when the data are Gaussian. As the percentage of the contaminating distribution in the mix increases, it becomes clear that  $r$  is the worst possible choice and  $r_p$  seems best. This paper will not dwell overlong on the comparison between  $r$ ,  $r_0$ , and  $r_p$ , since this has been done in great detail elsewhere (Wainer & Thissen, 1976). Rather, the results from  $r_p$ , the summary representation of a class of naive estimators, will be examined most carefully.

The correlation measure  $r_p$  seems reasonable when the parental correlation is low, although the other measures are at least as accurate and, as will be noted shortly, far more efficient. The statistic  $r_p$  is strikingly more accurate than the others when there is both a substantial correlation ( $\rho = .9$ ) and noise. Thus, on the basis of accuracy,  $r_p$  is preferred both when there is a strong relationship to be detected and when there is the possibility that this relationship is somewhat obscured by noise. Whenever there is no relationship,  $r_p$  does not discover one, but it is less efficient than the other measures.

A more careful examination of efficiency is therefore in order. In Table 2 are shown the relative efficiencies obtained from the sampling distribution of the four correlation statistics estimated directly from the monte carlo results. The number of observations per cell ( $n = 38$ ) is too small to estimate these accurately, but they do yield indications. (More precise estimates of the relative efficiencies of  $r$ ,  $r_0$ , and  $r_p$  are found in Wainer & Thissen, 1976.) The less accurate estimates used here provide a basis of comparison for the variance of  $r_p$ .

Table 1  
Monte Carlo 10% Trimmed Means for Four Estimators

	$\rho$	Percent Contamination	$r_p$	$r$	$r_t$	$r_0$
N=50	0.0	0	.1	.0	.0	.0
		10	.1	.0	.0	.0
		20	.1	.0	.0	.0
	0.5	0	.6	.5	.5	.5
		10	.5	.3	.4	.4
		20	.4	.3	.3	.3
	0.9	0	.9	.9	.9	.9
		10	.9	.6	.7	.7
		20	.8	.5	.6	.6
N=100	0.0	0	.1	.0	.0	.0
		10	.1	.0	.0	.0
		20	.0	.0	.0	.0
	0.5	0	.6	.5	.5	.5
		10	.4	.3	.3	.4
		20	.4	.3	.4	.3
	0.9	0	.9	.9	.9	.9
		10	.8	.6	.7	.8
		20	.8	.5	.6	.6

In any case, Table 2 shows rather clearly that the effects seen in the accuracy of  $r_p$  are mirrored in its variability. That is, when there is no underlying relationship,  $r_p$  has rather large variability and the other statistics are more reliable in their detection of nothing. However, when there is a substantial underlying linear structure,  $r_p$  actually becomes more efficient as noise increases relative to the other estimators. Although its efficiency is mediocre in any circumstances, in terms of accuracy with high correlation and substantial noise it is certainly far superior to the other estimators. The conclusion to be drawn is clear. When the standard estimators indicate that there is no relationship, this means either that there is no relationship or that a little contamination has effectively obscured it, and some estimators are worse than others. The authors' choice, for reasons to be elaborated upon shortly, is  $r_t$ ;  $r_p$ , on the other hand, will indicate a relationship with reasonable accuracy, when it is there, even when there is contamination. Thus, when looking for a relationship, use  $r_p$ ; and when looking for nothing, use any estimator desired, although the authors' preference is  $r_t$ .

Table 2  
Monte Carlo 10% Winzorized Relative Efficiencies for Four Estimators

	$\rho$	Percent Contamination	$r_p$	$r$	$r_t$	$r_0$
N=50	0.0	0	.5	1.0	1.0	.9
		10	.8	.9	.8	1.0
		20	.7	.7	1.0	.9
	0.5	0	.3	1.0	1.0	.9
		10	.2	.6	.9	1.0
		20	.4	.8	1.0	.7
	0.9	0	.5	1.0	.9	.5
		10	.4	.4	.7	1.0
		20	.3	.3	.5	1.0
N=100	0.0	0	.4	1.0	1.0	.9
		10	.6	.9	1.0	.7
		20	.3	.3	.6	1.0
	0.5	0	.2	.9	1.0	.4
		10	.2	.6	1.0	.8
		20	.2	.7	.7	1.0
	0.9	0	.1	1.0	1.0	.7
		10	.3	.5	1.0	.8
		20	.3	.4	.6	1.0

#### Estimation of $r_p$

The statistic designated  $r_p$  summarizes a class of naive estimators of correlation. To compute  $r_p$ , a class of naive estimators must first be found. A class of 38 graduate students who were taking an introductory course in statistics in the Department of Behavioral Sciences at The University of Chicago was used. They were each presented with different sets of nine scatter plots drawn from the design specified earlier and placed in a random order. At the front of the classroom were projected four scatterplots which were bivariate normal, having  $r$ 's of .0, .25, .5, and .75, respectively. The students were instructed to look at these as prototypes and then estimate the correlation in each of the scatterplots in their packet. They were not told to estimate the parental portion of the distribution, nor even that the distributions were made up of two parts. Each student had a different set of scatterplots, but all were random samples from the same overall design. This experiment was repeated for scatterplots with 50 points and with 100 points. The values of  $r_p$  (the "p" subscript stands for "people") reported were the trimmed means of these estimates.

The results obtained differ substantially from those reported by Strahan and Hansen (1978), who found that when the subjects did not have a prototype to work from and when the distributions were bivariate Gaussian, they tended, on the average, to underestimate the product-moment correlation. It was found in the present study, however, that there is no consistent underestimation of correlation when the data were Gaussian (contaminating distribution of 0%), which argues for the helpfulness of the prototype, although there was a positive bias for small (zero) correlations. The use of a trimmed mean to represent the subjects did not seem to affect the results. This estimator (subject's judgments) was not particularly efficient, which suggests that perhaps the sampling distribution of the subject's estimates may be fatter tailed than expected.

The more important aspect of this study is in the use of contaminated distributions. When data are bivariate Gaussian, anything will work. The more usual situation occurs when the researcher looks at a bivariate plot and finds that the "correlation" seen is quite different from the one calculated. The findings of the present study suggest that the researcher is more likely to be correct and that the statistic is adversely affected. Thus, it is suggested that there are a few more cells in the Strahan and Hansen design which should be worth being filled. The nature of  $r_p$  also explains why there were such an odd number of samples in each cell of the simulation.

### Conclusion and Discussion

The product-moment correlation was originally invented to characterize a particular form of data relationship. Its distribution theory is crucially based on the assumption of bivariate normality, and even small disturbances from this ideal may be disastrous. Thus, it can be seen that although  $r$  was developed to mirror intuition in some way, it fails seriously when the data are not as hypothesized. The obvious task is then to find another objective estimator that does correctly mirror intuition.

Toward this end, a number of alternatives have been proposed and tested. The rank-order correlation  $r_o$  is not a bad choice except that it has two problems: First, it is inefficient when there is a large underlying correlation; and second, it has no associated measure of scale with which to combine it into an integrated system of regression. However, if the absence of a relationship (the estimation of small correlations) is sought,  $r_o$  has much to recommend it. Tukey's  $r_t$  is a more versatile choice, though by no means perfect. It is very efficient relative to  $r$  when the data are normal and more accurate when they are not. Although for purposes of demonstration rather heavily contaminated distributions were used in this paper, it perhaps should be expected that in practice it would be expected that a 4% to 8% contamination would be found, in which case the Gaussian efficiency of  $r_t$  is important.

Additionally associated with  $r_t$  is a robust measure of scale,  $s^*$ . These can be combined to yield robust regression weights. Note, however, that  $r_t$  and  $r_o$  are still not as accurate as  $r_p$  when there is a strong relationship, indicating that a better estimator is possible. The accuracy of  $r_p$  gives ample support for proposing that it be used as an initial robust estimate of correlation for those adaptive procedures (Devlin, Gnanadesikan, & Kettenring, 1975) that require starting values for iterative methods.

The conclusion from this experiment in intuitive statistics is that graphical techniques are important and useful tools for the careful inspection of data and that if the statistics calculated disagree with intuition, it is likely that the statistics are wrong. Of course, this holds only for the kind of task in which the very powerful visual system is coupled with the appropriate mechanisms for pattern recognition. It most emphatically does not apply for those situations in which it is necessary to combine digital information in some optimal way. That task, as Meehl (1954), Dawes and Corrigan (1974), Einhorn and Hogarth (1975), and Wainer (1976) have all pointed out, is far better performed by re-



gression schemes than by people. Although Kahneman and Tversky (1973) have pointed out a number of instances in which human ability is disappointing, here at least is one instance in which no existing method can approach a class of naive estimators. The conclusion of this paper also hints that situations in which human decision making is poor (Dawes & Corrigan, 1974) might be substantially improved if more clever display techniques were used than merely tabulating a set of scores.

### References

- Box, G. E. P., & Tiao, G. C. A Bayesian approach to some outlier problems. *Biometrika*, 1968, *55*, 119-129.
- Dawes, R. M., & Corrigan, B. Linear models in decision making. *Psychological Bulletin*, 1974, *81*, 95-106.
- Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 1975, *62*, 531-546.
- Dixon, W. J. Simplified estimation from censored normal samples. *Annals of Mathematical Statistics*, 1960, *31*, 385-391.
- Downton, F. Linear estimates with polynomial coefficients. *Biometrika*, 1966, *53*, 129-141.
- Einhorn, H. J., & Hogarth, R. M. Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 1975, *13*, 171-192.
- Hogg, R. Adaptive robust procedures: A partial review and some suggestions for further applications and theory. *Journal of the American Statistical Association*, 1974, *69*, 909-927.
- Kahneman, D., & Tversky, A. On the psychology of prediction. *Psychological Review*, 1973, *80*, 237-251.
- Meehl, P. E. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press, 1954.
- Strahan, R. F., & Hansen, C. J. Underestimating correlation from scatterplots. *Applied Psychological Measurement*, 1978, *2*, 543-550.
- Tukey, J. W., & McLaughlin, D. H. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhya*, 1963, *A25*, 331-352.
- Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 1976, *83*, 213-217.
- Wainer, H., & Thissen, D. When Jackknifing fails (or does it?). *Psychometrika*, 1975, *40*, 113-114.
- Wainer, H., & Thissen, D. Three steps towards robust regression. *Psychometrika*, 1976, *41*, 9-34.

### Acknowledgments

*This research was supported in part by the National Science Foundation, Grant No. S0C76-17768 A01, to the Bureau of Social Science Research. Albert Biderman and Howard Wainer, project directors.*

### Author's Address

Send requests for reprints or further information to Howard Wainer, Bureau of Social Science Research, 1990 M Street, N.W., Suite 700, Washington, DC 20036.