Chapter 10
Correlation and Regression

Abstract This chapter introduces permutation methods for measures of correlation and regression, the best-known of which is Pearson's product-moment correlation coefficient. Included in this chapter are six example analyses illustrating computation of exact permutation probability values for correlation and regression, calculation of measures of effect size for measures of correlation and regression, the effects of extreme values on conventional (ordinary least squares) and permutation (least absolute deviation) correlation and regression, exact and Monte Carlo permutation procedures for measures of correlation and regression, application of permutation methods to correlation and regression with rank-score data, and analysis of multiple correlation and regression. Included in this chapter are permutation versions of ordinary least squares correlation and regression, least absolute deviation correlation and regression, Spearman's rank-order correlation coefficient, Kendall's rank-order correlation coefficient, Spearman's footrule measure of correlation, and a permutation-based alternative for the conventional measures of effect size for correlation and regression: Pearson's r^2.

This chapter presents exact and Monte Carlo permutation statistical methods for measures of linear correlation and regression. Also presented in this chapter is a permutation-based measure of effect size for a variety of measures of linear correlation and regression. Simple linear correlation coefficients between two variables constitute the foundation for a large family of advanced analytic techniques and are taught in every introductory course.

In this chapter, permutation statistical methods for measures of linear correlation and regression are illustrated with six example analyses. The first example utilizes a small set of observations to illustrate the computation of exact permutation methods for measures of linear correlation, wherein the permutation test statistic, δ, is developed and compared with Pearson's conventional product-moment correlation coefficient. The second example develops a permutation-based measure of effect size as a chance-corrected alternative to Pearson's squared product-moment correlation coefficient. The third example compares permutation statistical methods

K. J. Berry et al., *A Primer of Permutation Statistical Methods*,
https://doi.org/10.1007/978-3-030-20933-9_10

based on ordinary and squared Euclidean scaling functions, with an emphasis on the analysis of data sets containing extreme values. Ordinary least squares (OLS) regression, based on squared Euclidean scaling, and least absolute deviation (LAD) regression, based on ordinary Euclidean scaling, are compared and contrasted. The fourth example utilizes a larger data set for providing comparisons of exact permutation methods and Monte Carlo permutation methods, demonstrating the efficiency of Monte Carlo statistical methods for correlation analyses. The fifth example illustrates the application of permutation statistical methods to univariate rank-score data, comparing permutation statistical methods with Spearman's rank-order correlation coefficient, Kendall's rank-order correlation coefficient, and Spearman's footrule measure of rank-order correlation. The sixth example illustrates the application of permutation statistical methods to multivariate correlation and regression. Both OLS and LAD multivariate linear regression are described and compared for multivariate observations.

10.1 Introduction

The most popular measure of linear correlation between two interval-level variables, say, x and y, is Pearson's r_{xy} product-moment correlation coefficient wherein the Neyman–Pearson null hypothesis (H_0) posits a value for a population parameter, such as a population correlation coefficient; that is, H_0: $\rho_{xy} = \theta$, where θ is a specified value between -1 and $+1$. For example, the null hypothesis might stipulate that the correlation in the population from which a bivariate sample has been drawn is H_0: $\rho_{xy} = 0$. In this chapter the null hypothesis, H_0: $\rho_{xy} = 0$, is used exclusively for two reasons. First, most introductory courses in statistical methods restrict their discussions to H_0: $\rho_{xy} = 0$. Null hypotheses such as H_0: $\rho_{xy} \neq 0$ are usually treated in more advanced courses. Second, Fisher's normalizing transformation for r_{xy} when $\rho_{xy} \neq 0$ has been found to be unsatisfactory unless either the population correlation coefficient $\rho_{xy} = 0$ or the population is known to be bivariate normal [4].

The problem is easy to illustrate. Consider a population in which the product-moment correlation is equal to zero; that is, $\rho_{xy} = 0$, such as depicted in Fig. 10.1. Random sampling from a population in which $\rho_{xy} = 0$ produces a symmetric, discrete sampling distribution of r_{xy} values that can be approximated by Student's t distribution with $N - 2$ degrees of freedom, such as depicted in Fig. 10.2.

Now consider a population in which the product-moment correlation is not equal to zero; that is, $\rho_{xy} = +0.60$, such as depicted in Fig. 10.3. Random sampling from a population in which $\rho_{xy} = +0.60$ produces an negatively-skewed, discrete sampling distribution of r_{xy} values that cannot be approximated by Student's t distribution with $N - 2$ degrees of freedom, such as depicted in Fig. 10.4.

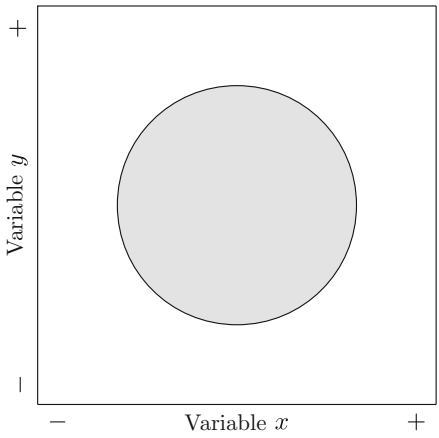

Fig. 10.1 Simulated scatterplot of a population with $\rho_{xy} = 0.00$

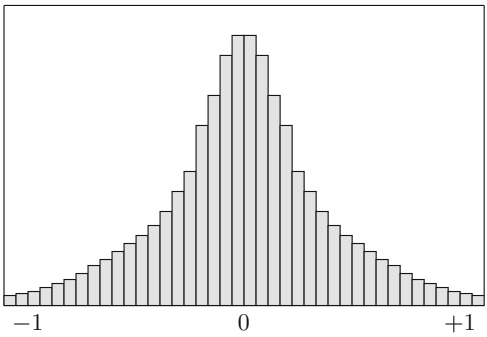

Fig. 10.2 Simulated discrete permutation distribution of r_{xy} from a population with $\rho_{xy} = +0.00$

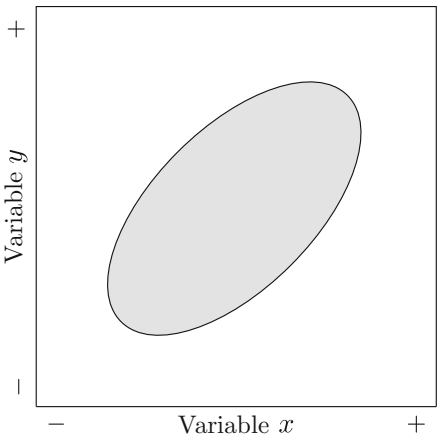

Fig. 10.3 Simulated scatterplot of a population with $\rho_{xy} = +0.60$

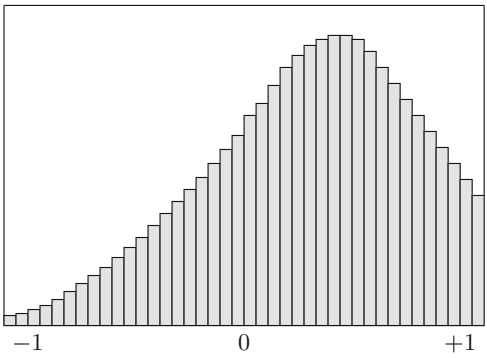

Fig. 10.4 Simulated discrete permutation distribution of r_{xy} from a population with $\rho_{xy} = +0.60$

For simple linear correlation with two interval-level variables and N paired observations, Pearson's product-moment correlation coefficient is given by

$$r_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N} (y_i - \bar{y})^2\right]}} ,$$

where $\bar{x}$ and $\bar{y}$ denote the arithmetic means of variables x and y given by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i ,$$

respectively, x_i and y_i denote the ith observed sample values for $i = 1, \ldots, N$, and N is the number of bivariate observations.

Under the Neyman–Pearson population model the null hypothesis is H_0: $\rho_{xy} = \theta$ and the two-tail alternative hypothesis is H_1: $\rho_{xy} \neq \theta$, where θ is a hypothesized value for the population correlation coefficient. The conventional test of significance for Pearson's product-moment correlation coefficient with null hypothesis, H_0: $\rho_{xy} = 0$, is Student's t test statistic given by

$$t = r_{xy}\sqrt{\frac{N-2}{1-r_{xy}^2}} ,$$

which is assumed to follow Student's t distribution with $N - 2$ degrees of freedom, under the assumptions of normality and homogeneity. The permissible probability of a type I error is denoted by α and if the observed value of t is more extreme than the critical values of $\pm t$ that define α, the null hypothesis is rejected with a probability of type I error equal to or less than α. The test of significance does

not determine whether or not the null hypothesis is true, but only provides the probability that, if the null hypothesis is true, the sample has been drawn from a population with the value specified under the null hypothesis.

The assumptions underlying Pearson's product-moment correlation coefficient are (1) the observations are independent, (2) the data are a random sample from a well-defined population with $\rho_{xy} = 0$, (3) the relationship between the predictor variable and the criterion variable is linear, (4) homogeneity of variance, and (5) the target variables x and y are distributed bivariate normal in the population.

10.1.1 A Permutation Approach

Consider a simple linear correlation analysis between two variables under the Fisher–Pitman permutation model of statistical inference. As discussed in previous chapters, the permutation model differs from the Neyman–Pearson population model in several ways. Under the Fisher–Pitman permutation model there is no null hypothesis specifying a population parameter. Instead, the Fisher–Pitman null hypothesis simply states that all possible arrangements of the observed data occur with equal chance [5]. Also, there is no alternative hypothesis under the permutation model and no specified α level. Moreover, there is no requirement of random sampling, no degrees of freedom, and no assumption of normality or homogeneity. Finally, the Fisher–Pitman permutation statistical model provides exact probability values.

A permutation alternative to a conventional correlation analysis for two variables is easily defined. Let x_i and y_i denote the paired sample values for $i = 1, \ldots, N$. The permutation test statistic is given by

$$\delta = S_x^2 + S_y^2 - 2|r_{xy}|S_x S_y + (\bar{x} - \bar{y})^2 ,$$

where the sample means for variables x and y are given by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i ,$$

respectively, and the sample variances for variables x and y are given by

$$S_x^2 = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2 \quad \text{and} \quad S_y^2 = \frac{1}{N}\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2 ,$$

respectively.[1]

[1]Note that whereas a permutation approach eschews estimated population parameters and degrees of freedom, the summations are divided by N, not $N - 1$. Thus S_x^2 and S_y^2 denote the sample variances, not the estimated population variances.

Under the Fisher–Pitman null hypothesis, the exact probability value of an observed δ is the proportion of δ test statistic values calculated on all possible arrangements of the observed data that are equal to or less than the observed value of δ; that is,

$$P(\delta \leq \delta_{\mathrm{o}} | H_0) = \frac{\text{number of } \delta \text{ values } \leq \delta_{\mathrm{o}}}{M} ,$$

where δ_{o} denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements in the reference set of all permutations of the observed data.

10.2 Example 1: The Relationship Between r_{xy} and δ

An example will serve to illustrate the relationships between test statistics r_{xy} and δ for a simple correlation analysis. Consider the small set of data listed in Table 10.1 with $N = 4$ bivariate observations. For the example bivariate observations listed in Table 10.1, the sample means for variables x and y are

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{24 + 31 + 55 + 43}{4} = 38.25$$

and

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{20 + 36 + 49 + 35}{4} = 35.00 ,$$

respectively, the sample product-moment correlation coefficient is

$$r_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N} (y_i - \bar{y})^2\right]}}$$

$$= \frac{+441.00}{\sqrt{(558.75)(422.00)}} = +0.9082 ,$$

Table 10.1 Example correlation data on $N = 4$ bivariate observations

	Variable	
Object	x	y
1	24	20
2	31	36
3	55	49
4	43	35

and Student's t test statistic is

$$t = r_{xy}\sqrt{\frac{N-2}{1-r_{xy}^2}} = +0.9082\sqrt{\frac{4-2}{1-(+0.9082)^2}} = +3.0684\,.$$

Under the Neyman–Pearson null hypothesis, H_0: $\rho_{xy} = 0$, test statistic t is asymptotically distributed as Student's t with $N-2$ degrees of freedom. With $N-2 = 4-2 = 2$ degrees of freedom, the asymptotic two-tail probability value of $t = +3.0684$ is $P = 0.0918$, under the assumptions of linearity, normality, and homogeneity.

10.2.1 An Exact Permutation Analysis

Now consider the bivariate data listed in Table 10.1 under the Fisher–Pitman permutation model. For the example bivariate data listed in Table 10.1, the sample means are $\bar{x} = 38.25$ and $\bar{y} = 35.00$, the sample variances are $S_x^2 = 139.6875$ and $S_y^2 = 105.50$, the sample standard deviations are $S_x = 11.8189$ and $S_y = 10.2713$, the sample product-moment correlation coefficient is $r_{xy} = +0.9082$, and the observed permutation test statistic is

$$\begin{aligned}\delta &= S_x^2 + S_y^2 - 2|r_{xy}|S_xS_y + (\bar{x}-\bar{y})^2 = 139.6875 + 105.50 \\ &\quad - 2(0.9082)(11.8189)(10.2713) + (38.25-35.00)^2 = 35.25\,. \qquad (10.1)\end{aligned}$$

Note that in Eq. (10.1), S_x, S_x^2, S_y, S_y^2, $\bar{x}$, $\bar{y}$, and the constant 2 are all invariant under permutation, leaving only $|r_{xy}|$ to be calculated for each arrangement of the observed data.

An exact permutation analysis requires exhaustive shuffles of either the $N = 4$ x values or the $N = 4$ y values while holding the other set of values constant. For the example data listed in Table 10.1 there are only

$$M = N! = 4! = 24$$

possible, equally-likely arrangements in the reference set of all permutations of the bivariate data listed in Table 10.1, making an exact permutation analysis feasible. Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the $N = 4$ bivariate observations listed in Table 10.1 that are equal to or less than the observed value of $\delta = 35.25$. Table 10.2 lists the $M = 24$ arrangements of the example data listed in Table 10.1 with the x values shuffled and the associated values for r_{xy} and δ, ordered by the $|r_{xy}|$ values from largest ($|r_1| = 0.9432$) to smallest ($|r_{24}| = 0.1524$) and by the δ values from smallest

Table 10.2 All $M = 24$ possible, equally-likely arrangements of the bivariate data listed in Table 10.1

Arrangement	Variable x	Variable y	$\lvert r_{xy} \rvert$	δ
1*	55, 31, 24, 43	20, 36, 49, 35	0.9432	26.75
2*	24, 43, 55, 31	20, 36, 49, 35	0.9329	29.25
3*	55, 43, 24, 31	20, 36, 49, 35	0.9185	32.75
4*	24, 31, 55, 43	20, 36, 49, 35	0.9082	35.25
5	55, 24, 31, 43	20, 36, 49, 35	0.7558	72.25
6	31, 43, 55, 24	20, 36, 49, 35	0.7167	81.75
7	55, 43, 31, 24	20, 36, 49, 35	0.7167	81.75
8	31, 24, 55, 43	20, 36, 49, 35	0.6775	91.25
9	24, 55, 43, 31	20, 36, 49, 35	0.6116	107.25
10	43, 31, 24, 55	20, 36, 49, 35	0.5725	116.75
11	24, 31, 43, 55	20, 36, 49, 35	0.5622	119.25
12	43, 55, 24, 31	20, 36, 49, 35	0.5231	128.75
13	55, 24, 43, 31	20, 36, 49, 35	0.4098	156.25
14	31, 55, 43, 24	20, 36, 49, 35	0.3954	159.75
15	55, 31, 43, 24	20, 36, 49, 35	0.3954	159.75
16	43, 24, 31, 55	20, 36, 49, 35	0.3851	162.25
17	31, 24, 43, 55	20, 36, 49, 35	0.3316	175.25
18	43, 31, 55, 24	20, 36, 49, 35	0.3213	177.75
19	43, 55, 31, 24	20, 36, 49, 35	0.3213	177.75
20	43, 24, 55, 31	20, 36, 49, 35	0.3068	181.25
21	24, 55, 31, 43	20, 36, 49, 35	0.2657	191.25
22	24, 43, 31, 55	20, 36, 49, 35	0.2409	197.25
23	31, 43, 24, 55	20, 36, 49, 35	0.1771	212.75
24	31, 55, 24, 43	20, 36, 49, 35	0.1524	218.75

($\delta_1 = 29.25$) to largest ($\delta_{24} = 218.75$). For test statistic δ there are four δ test statistic values that are equal to or less than the observed value of $\delta = 35.25$ ($\delta_1 = 26.75, \delta_2 = 29.25, \delta_3 = 32.75$, and $\delta_4 = 35.25$). The arrangements yielding the four smallest δ values are indicated with asterisks in Table 10.2. If all M arrangements of the $N = 4$ bivariate observations listed in Table 10.1 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 35.25$ computed on the $M = 24$ possible arrangements of the observed data with $N = 4$ bivariate observations preserved for each arrangement is

$$P(\delta \leq \delta_{\mathrm{o}}) = \frac{\text{number of } \delta \text{ values } \leq \delta_{\mathrm{o}}}{M} = \frac{4}{24} = 0.1667 \,,$$

where δ_{o} denotes the observed value of δ and M is the number of possible, equally-likely arrangements of the $N = 4$ bivariate observations listed in Table 10.1.

Alternatively, since test statistics δ and r_{xy} are equivalent under the Fisher–Pitman null hypothesis, there are four $|r_{xy}|$ values that are equal to or greater than

the observed value of $|r_{xy}| = 0.9082$ ($|r_1| = 0.9432$, $|r_2| = 0.9329$, $|r_3| = 0.9185$, and $|r_4| = 0.9082$) yielding an exact probability value for $|r_{xy}| = 0.9082$ of

$$P(|r_{xy}| \geq |r_o|) = \frac{\text{number of } |r_{xy}| \text{ values } \geq |r_o|}{M} = \frac{4}{24} = 0.1667 ,$$

where $|r_o|$ denotes the observed value of $|r_{xy}|$. There is a considerable difference between the asymptotic probability value for r_{xy} based on Student's t distribution ($P = 0.0918$) and the exact permutation probability value for δ ($P = 0.1667$). The actual difference between the two probability values is

$$\Delta_P = 0.1667 - 0.0918 = 0.0749 .$$

The difference is most probably due to the very small number of arrangements of the observed data. A continuous mathematical function such as Student's t cannot be expected to provide a precise fit to only 24 data points of which only 21 are different.

10.3 Example 2: Measures of Effect Size

Measures of effect size express the practical or clinical significance of a sample correlation coefficient, as contrasted with the statistical significance of the correlation coefficient. For an illustration of the measurement of effect size, consider the example data listed in Table 10.3 with $N = 11$ bivariate observations. The standard measure of effect size is simply the squared Pearson product-moment correlation between variables x and y. For the example bivariate data listed in Table 10.3, the

Table 10.3 Example correlation data on $N = 11$ bivariate observations

Object	Variable	
	x	y
1	11	4
2	18	11
3	12	1
4	27	16
5	15	5
6	21	9
7	25	10
8	15	2
9	18	8
10	23	7
11	12	3

sample means for variables x and y are

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{11 + 18 + \cdots + 12}{11} = 17.9091$$

and

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{4 + 11 + \cdots + 3}{11} = 6.9091 ,$$

respectively, the sample product-moment correlation coefficient is

$$r_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N} (y_i - \bar{y})^2\right]}}$$

$$= \frac{+209.9091}{\sqrt{(302.9091)(200.9091)}} = +0.8509 ,$$

the squared product-moment measure of effect size is

$$r_{xy}^2 = (+0.8509)^2 = 0.7240 ,$$

and Student's t test statistic is

$$t = r_{xy}\sqrt{\frac{N-2}{1-r_{xy}^2}} = +0.8509\sqrt{\frac{11-2}{1-(+0.8509)^2}} = +4.8592 .$$

Under the Neyman–Pearson null hypothesis, H_0: $\rho_{xy} = 0$, test statistic t is asymptotically distributed as Student's t with $N - 2$ degrees of freedom. With $N - 2 = 11 - 2 = 9$ degrees of freedom, the asymptotic two-tail probability value of $t = +4.8592$ is $P = 0.8969 \times 10^{-3}$, under the assumptions of linearity, normality, and homogeneity.

10.3.1 An Exact Permutation Analysis

Now consider the example data listed in Table 10.3 under the Fisher–Pitman permutation model. For the example data listed in Table 10.3, the sample means

are $\bar{x} = 17.9091$ and $\bar{y} = 6.9091$, the sample variances are $S_x^2 = 27.5372$ and $S_y^2 = 18.2645$, the sample standard deviations are $S_x = 5.2476$ and $S_y = 4.2737$, the sample product-moment correlation coefficient is $r_{xy} = +0.8509$, and the observed permutation test statistic is

$$\delta = S_x^2 + S_y^2 - 2|r_{xy}|S_x S_y + (\bar{x} - \bar{y})^2 = 27.5372 + 18.2645$$
$$- 2(0.8509)(5.2476)(4.2737) + (17.9091 - 6.9091)^2 = 128.6364 .$$

An exact permutation analysis requires shuffling of either the $N = 11$ x values or the $N = 11$ y values while holding the other set of values constant. For the example data listed in Table 10.3 there are

$$M = N! = 11! = 39{,}916{,}800$$

possible, equally-likely arrangements in the reference set of all permutations of the observed bivariate data, making an exact permutation analysis feasible.

The exact expected value of the $M = 39{,}916{,}800$ δ test statistic values under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M} \delta_i = \frac{6{,}658{,}188{,}218}{39{,}916{,}800} = 166.8017 .$$

Alternatively, the exact expected value of test statistic δ is

$$\mu_\delta = S_x^2 + S_y^2 + (\bar{x} - \bar{y})^2$$
$$= 27.5372 + 18.2645 + (17.9091 - 6.9091)^2 = 166.8017 .$$

The observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{128.6364}{166.8017} = +0.2288 ,$$

indicating approximately 23% agreement between the x and y values above what is expected by chance.

Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the $N = 11$ bivariate observations that are equal to or less than the observed value of $\delta = 128.6364$. There are exactly 35,216 δ test statistic values that are equal to or less than the observed value of $\delta = 128.6364$. If all M arrangements of the $N = 11$ bivariate observations listed in Table 10.3 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 128.6364$ computed on the $M = 39{,}916{,}800$ possible arrangements

of the observed data with $N = 11$ bivariate observations preserved for each arrangement is

$$P(\delta \le \delta_o | H_0) = \frac{\text{number of } \delta \text{ values } \le \delta_o}{M} = \frac{35{,}216}{39{,}916{,}800} = 0.8822 \times 10^{-3} ,$$

where δ_o denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 11$ bivariate observations listed in Table 10.3. In this example there are 39,916,800 data points to be fit by Student's t distribution and there are no extreme values. Thus the asymptotic probability value ($P = 0.8969 \times 10^{-3}$) and the exact permutation probability value ($P = 0.8822 \times 10^{-3}$) are similar, with a difference between the probability values of only

$$\Delta_P = 0.8969 \times 10^{-3} - 0.8822 \times 10^{-3} = 0.1459 \times 10^{-4} .$$

10.4 Example 3: Analyses with $v = 2$ and $v = 1$

Ordinary least squares (OLS) linear regression and correlation have long been recognized as useful tools in many areas of research. The optimal properties of OLS linear regression and correlation are well known when the errors are normally distributed. However, in practice the assumption of normality is rarely justified. Least absolute deviation (LAD) regression and correlation are often superior to OLS linear regression and correlation when the errors are not normally distributed. Estimators of OLS regression parameters can be severely affected by unusual values in the criterion variable, in one or more of the predictor variables, or both. In contrast, LAD regression is less sensitive to the effects of unusual variables because the errors are not squared [3]. The effect of extreme values on OLS and LAD regression and correlation is analogous to the effect of extreme values on the mean and median as measures of location.

Consider N paired x_i and y_i observed values for $i = 1, \ldots, N$. For the OLS regression equation given by

$$\hat{y}_i = \hat{\alpha}_{yx} + \hat{\beta}_{yx} x_i ,$$

where $\hat{y}_i$ is the ith of N predicted criterion values and x_i is the ith of N predictor values, $\hat{\alpha}_{yx}$ and $\hat{\beta}_{ys}$ are the OLS parameter estimators of the population intercept (α_{yx}) and population slope (β_{yx}), respectively, and are given by

$$\hat{\beta}_{yx} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

and

$$\hat{\alpha}_{yx} = \bar{y} - \hat{\beta}_{yx}\bar{x} \ ,$$

where $\bar{x}$ and $\bar{y}$ are the sample means of variables x and y, respectively. Estimators of OLS regression parameters minimize the sum of the squared differences between the observed (y_i) and predicted ($\hat{y}_i$) criterion values for $i = 1, \ldots, N$; that is,

$$\sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|^v \ ,$$

where for OLS regression based on a squared Euclidean scaling function, $v = 2$.

For the LAD regression equation given by

$$\tilde{y}_i = \tilde{\alpha}_{yx} + \tilde{\beta}_{yx} x_i \ ,$$

where $\tilde{y}_i$ is the ith of N predicted criterion values and x_i is the ith of N predictor values, $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ are the LAD parameter estimators of the population intercept (α_{yx}) and population slope (β_{yx}), respectively.[2]

Unlike OLS regression, no simple expressions can be given for LAD regression estimators $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$. However, values for $\tilde{\alpha}_{yx}$ and $\tilde{\beta}_{yx}$ may be found through an efficient linear programming algorithm, such as provided by Barrodale and Roberts [1, 2]. In contrast to estimators of OLS regression parameters, estimators of LAD regression parameters minimize the sum of the absolute differences between the observed (y_i) and predicted ($\tilde{y}_i$) criterion values for $i = 1, \ldots, N$; that is,

$$\sum_{i=1}^{N} \left| y_i - \tilde{y}_i \right|^v \ ,$$

where for LAD regression based on ordinary Euclidean scaling, $v = 1$.

For LAD regression it is convenient to have a measure of agreement, not product-moment correlation, between the observed and predicted y values. Let the permutation test statistic be given by

$$\delta = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \tilde{y}_i \right|^v$$

[2]In this section, a caret(∧) over a symbol such as $\hat{\alpha}$ or $\hat{\beta}$ indicates an OLS regression model predicted value of a corresponding population parameter, while a tilde (~) over a symbol such as $\tilde{\alpha}$ or $\tilde{\beta}$ indicates a LAD regression model predicted value of a corresponding population parameter.

Table 10.4 Example bivariate correlation data on $N = 10$ subjects

Subject	x	y
1	14	25
2	8	23
3	5	21
4	2	10
5	1	12
6	3	11
7	9	19
8	2	13
9	3	13
10	9	16

and let $v = 1$ for correspondence with LAD regression. Then the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis is given by

$$\mu_\delta = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| y_i - \tilde{y}_j \right|^v ,$$

and a chance-corrected measure of agreement between the observed y_i values and the LAD predicted $\tilde{y}_i$ values for $i = 1, \ldots, N$ is given by

$$\Re = 1 - \frac{\delta}{\mu_\delta} .$$

10.4.1 An Example OLS Regression Analysis

To illustrate the relative differences between OLS and LAD regression, consider the small example set of bivariate values listed in Table 10.4 for $N = 10$ subjects. For the bivariate data listed in Table 10.4 the OLS estimate of the population slope is $\hat{\beta}_{yx} = +1.0673$, the OLS estimate of the population intercept is $\hat{\alpha}_{yx} = +10.3229$, and the Pearson product-moment correlation coefficient is $r_{xy} = +0.8414$. Table 10.5 lists the $N = 10$ observed values for variables x and y, the OLS predicted y values ($\hat{y}$), the residual errors ($\hat{e}$), and the squared residual errors ($\hat{e}^2$). Under the Neyman–Pearson population model Pearson's product-moment correlation coefficient is asymptotically distributed as Student's t under the null hypothesis, H_0: $\rho_{xy} = 0$, with $N - 2$ degrees of freedom.[3] For the $N = 10$ bivariate

[3] One degree of freedom is lost due to the sample estimate ($\hat{\alpha}_{yx}$) of the population intercept and one degree of freedom is lost due to the sample estimate ($\hat{\beta}_{yx}$) of the population slope.

Table 10.5 Observed x and y values with associated predicted values ($\hat{y}$), residual errors ($\hat{e}$), and squared residual errors ($\hat{e}^2$) from the bivariate correlation data listed in Table 10.4

Subject	x	y	$\hat{y}$	$\hat{e}$	$\hat{e}^2$
1	14	25	25.2656	−0.2656	0.0705
2	8	23	18.8616	+4.1384	17.1264
3	5	21	15.6596	+5.3404	28.5199
4	2	10	12.4576	−2.4576	6.0398
5	1	12	11.3903	+0.6097	0.3718
6	3	11	13.5249	−2.5249	6.3753
7	9	19	19.9289	−0.9289	0.8629
8	2	13	12.4576	+0.5424	0.2942
9	3	13	13.5249	−0.5249	0.2756
10	9	16	19.9289	−3.9289	15.4365
Sum	56	163	163.0000	0.0000	75.3728

observations listed in Table 10.4 with $N - 2 = 10 - 2 = 8$ degrees of freedom Student's test statistic,

$$t = r_{xy}\sqrt{\frac{N-2}{1-r_{xy}^2}} = +0.8414\sqrt{\frac{10-2}{1-(+0.8414)^2}} = +4.4039\,,$$

yields an asymptotic two-tail probability value of $P = 0.2275\times10^{-2}$, under the assumptions of linearity, normality, and homogeneity.

10.4.2 An Example LAD Regression Analysis

For the bivariate data listed in Table 10.4, the LAD estimate of the population intercept is $\tilde{\alpha}_{yx} = +9.7273$, the LAD estimate of the population slope is $\tilde{\beta}_{yx} = +1.0909$, the observed permutation test statistic is

$$\delta = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \tilde{y}_i\right| = \frac{20.6364}{10} = 2.0636\,,$$

the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|y_i - \tilde{y}_j\right| = \frac{533.8182}{10^2} = 5.3382\,,$$

Table 10.6 Observed x and y values with associated predicted values ($\tilde{y}$), residual errors ($\tilde{e}$), and absolute residual errors ($|\tilde{e}|$) from the bivariate correlation data listed in Table 10.4

Subject	x	y	$\tilde{y}$	$\tilde{e}$	$\lvert\tilde{e}\rvert$
1	14	25	25.0000	0.0000	0.0000
2	8	23	18.4545	+4.5455	4.5455
3	5	21	15.1818	+5.8182	5.8182
4	2	10	11.9090	−1.9091	1.9091
5	1	12	10.8182	+1.1818	1.1818
6	3	11	13.0000	−2.0000	2.0000
7	9	19	19.5455	−0.5455	0.5455
8	2	13	11.9090	+1.0909	1.0909
9	3	13	13.0000	0.0000	0.0000
10	9	16	19.5455	−3.5455	3.5455
Sum	56	163	158.3636	+4.6364	20.6364

and the chance-corrected measure of agreement between the observed y values and the LAD predicted $\tilde{y}$ values is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{2.0636}{5.3382} = +0.6134\,,$$

indicating approximately 61% agreement between the observed and predicted values of variable y.

Table 10.6 lists the $N = 10$ observed values of variables x and y, the predicted y values ($\tilde{y}$), the residual errors ($\tilde{e}$), and the absolute residual errors ($|\tilde{e}|$).

Since there are only

$$M = N! = 10! = 3{,}628{,}800$$

possible, equally-likely arrangements in the reference set of all permutations of the bivariate data listed in Table 10.4, an exact permutation analysis is possible. Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the $N = 10$ bivariate observations listed in Table 10.4 that are equal to or less than the observed value of $\delta = 2.0636$. Alternatively, the exact probability value of an observed $\Re$ agreement coefficient is the proportion of $\Re$ values computed on all possible, equally-likely arrangements of the $N = 10$ bivariate observations listed in Table 10.4 that are equal to or greater than the observed value of $\Re = +0.6134$. There are exactly 15,533 $\Re$ test statistic values that are equal to or greater than the observed value of $\Re = +0.6134$.

If all M arrangements of the $N = 10$ bivariate observations listed in Table 10.4 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\Re = +0.6134$ computed on the $M = 3{,}628{,}800$ possible arrangements of the observed data with $N = 10$ bivariate observations preserved

for each arrangement is

$$P(\Re \geq \Re_o | H_0) = \frac{\text{number of } \Re \text{ values } \geq \Re_o}{M} = \frac{15{,}533}{3{,}628{,}800} = 0.4280 \times 10^{-2} ,$$

where $\Re_o$ denotes the observed value of test statistic $\Re$.

Alternatively, since $\mu_\delta = 5.3382$ is a constant,

$$P(\delta \leq \delta_o | H_0) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{M} = \frac{15{,}533}{3{,}628{,}800} = 0.4280 \times 10^{-2} ,$$

where δ_o denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 10$ bivariate observations listed in Table 10.4.

10.4.3 The Effects of Extreme Values

For the example bivariate data listed in Table 10.4 on p. 374, the exact probability value based on LAD regression and ordinary Euclidean scaling with $v = 1$ is $P = 0.4280 \times 10^{-2}$ and the asymptotic probability value based on OLS regression and squared Euclidean scaling with $v = 2$ is $P = 0.2275 \times 10^{-2}$. In this case the difference between the asymptotic and exact probability values is only

$$\Delta_P = 0.4280 \times 10^{-2} - 0.2275 \times 10^{-2} = 0.2006 \times 10^{-2} .$$

The small difference in probability values is due to the fact that there are no extreme values in the data listed in Table 10.4 on p. 374. OLS analyses based on squared Euclidean scaling with $v = 2$ are mean-based and LAD analyses based on ordinary Euclidean scaling with $v = 1$ are median-based. Consequently, LAD regression analyses are highly resistant to extreme values.

Extreme values are common in applied research. To demonstrate the difference between OLS analyses based on squared Euclidean scaling with $v = 2$ and LAD analyses based on ordinary Euclidean scaling with $v = 1$ when the data contain an extreme value, consider the bivariate data listed in Table 10.7. The data listed in Table 10.7 are the same data listed in Table 10.4 on p. 374 with one alteration: the value of $y_2 = 23$ has been increased to $y_2 = 90$, thereby providing an extreme value.

For the bivariate data listed in Table 10.4 on p. 374 without an extreme value ($y_2 = 23$), the OLS sample correlation coefficient is $r_{xy} = +0.8414$, Student's t test statistic is $t = +4.4039$, and the asymptotic probability value to six decimal places is $P = 0.002275$. For the bivariate data listed in Table 10.7 with an extreme value ($y_2 = 90$), the OLS sample correlation coefficient is $r_{xy} = +0.3636$, Student's t test statistic is $t = +1.1042$, and the asymptotic probability value is $P = 0.301606$.

Table 10.7 Example bivariate LAD correlation data on $N = 10$ subjects with an extreme value included

Subject	x	y
1	14	25
2	8	90
3	5	21
4	2	10
5	1	12
6	3	11
7	9	19
8	2	13
9	3	13
10	9	16

The difference between the two OLS correlation coefficients is

$$\Delta_{r_{xy}} = 0.8414 - 0.3636 = 0.4778$$

and the difference between the two OLS probability values is

$$\Delta_P = 0.301606 - 0.002275 = 0.299331 .$$

For the bivariate data listed in Table 10.4 on p. 374 without an extreme value ($y_2 = 23$), the LAD agreement measure is $\Re = +0.6134$ and the exact probability value to six decimal places is $P = 0.004280$. For the bivariate data listed in Table 10.7 with an extreme value ($y_2 = 90$), the LAD agreement measure is $\Re = +0.2696$ and the exact probability value is $P = 0.006317$. The difference between the two LAD agreement measures is

$$\Delta_{\Re} = 0.6134 - 0.2696 = 0.3438$$

and the difference between the two LAD probability values is

$$\Delta_P = 0.006317 - 0.004280 = 0.002037 .$$

The difference between the two LAD agreement measures ($\Delta_{\Re} = 0.3438$) is considerably smaller than the difference between the two OLS correlation coefficients ($\Delta_{r_{xy}} = 0.4778$) and the difference between the two LAD probability values ($\Delta_P = 0.002037$) is almost two orders of magnitude smaller than the difference between the two OLS probability values ($\Delta_P = 0.299331$). While the LAD regression analysis of the data listed in Table 10.7 is clearly affected by the presence of an extreme value, LAD regression based on ordinary Euclidean scaling with $v = 1$ is a robust procedure relative to OLS regression based on squared Euclidean scaling with $v = 2$ when extreme values are present.

10.5 Example 4: Exact and Monte Carlo Analyses

As sample sizes become large, the number of possible arrangements of the observed data makes exact permutation methods impractical. For example, for a sample size of $N = 20$ there are

$$M = N! = 20! = 2{,}432{,}902{,}008{,}176{,}640{,}000$$

possible, equally-likely arrangements in the reference set of all permutations of the observed data to be analyzed. Far too many arrangements to be practical. Monte Carlo permutation methods examine a random sample of all M possible arrangements of the observed data, providing efficient and accurate results. Provided that the probability value is not too small, $L = 1{,}000{,}000$ random arrangements are usually sufficient to ensure three decimal places of accuracy [6].

For a fourth, larger example of bivariate correlation, consider the data on $N = 12$ objects listed in Table 10.8 under the Neyman–Pearson population model. For the example data listed in Table 10.8 with $N = 12$ bivariate observations, the means of variables x and y are

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{9 + 10 + \cdots + 8}{12} = 17.3333$$

and

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{21 + 25 + \cdots + 18}{12} = 6.9167\,,$$

Table 10.8 Example correlation data on $N = 12$ bivariate observations

Object	Variable	
	x	y
1	9	21
2	10	25
3	2	15
4	4	11
5	5	15
6	16	27
7	1	12
8	11	18
9	7	11
10	3	12
11	7	23
12	8	18

respectively, and Pearson's product-moment correlation coefficient between variables x and y is

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N}(y_i - \bar{y})^2\right]}}$$

$$= \frac{+209.3333}{\sqrt{\left[(346.6667)(200.9167)\right]}} = +0.7932 .$$

The conventional test of significance for Pearson's product-moment correlation coefficient is given by

$$t = r_{xy}\sqrt{\frac{N-2}{1-r_{xy}^2}} .$$

Under the Neyman–Pearson null hypothesis, H_0: $\rho_{xy} = 0$, test statistic t is asymptotically distributed as Student's t with $N - 2$ degrees of freedom.

For the example data listed in Table 10.8,

$$t = r_{xy}\sqrt{\frac{N-2}{1-r_{xy}^2}} = +0.7932\sqrt{\frac{12-2}{1-(+0.7932)^2}} = +4.1188$$

and with $N - 2 = 12 - 2 = 10$ degrees of freedom the asymptotic two-tail probability value is $P = 0.2081 \times 10^{-2}$, under the assumptions of linearity, normality, and homogeneity.

10.5.1 A Monte Carlo Permutation Analysis

Now consider the data listed in Table 10.8 under the Fisher–Pitman permutation model. For the example data listed in Table 10.8 with $N = 12$ bivariate observations, the sample means are $\bar{x} = 6.9167$ and $\bar{y} = 17.3333$, the sample variances are $S_x^2 = 16.7431$ and $S_y^2 = 28.8889$, the sample standard deviations are $S_x = 4.0918$ and $S_y = 5.3748$, the sample product-moment correlation coefficient is $r_{xy} = +0.7932$, and the observed permutation test statistic is

$$\delta = S_x^2 + S_y^2 - 2|r_{xy}|S_x S_y + (\bar{x} - \bar{y})^2 = 16.7431 + 28.8889$$

$$- 2(0.7932)(4.0918)(5.3748) + (6.9167 - 17.3333)^2 = 119.25 .$$

A permutation analysis of correlation requires shuffling either the $N = 12$ x values or the $N = 12$ y values, while holding the other variable constant. Even with the small sample of $N = 12$ bivariate observations, there are

$$M = N! = 12! = 479{,}001{,}600$$

possible, equally-likely arrangements in the reference set of all permutations of the example data listed in Table 10.8, making an exact permutation analysis impractical. Under the Fisher–Pitman permutation model, the Monte Carlo probability of an observed δ is the proportion of δ test statistic values computed on the randomly-selected, equally-likely arrangements of the $N = 12$ bivariate observations listed in Table 10.8 that are equal to or less than the observed value of $\delta = 119.25$. Based on $L = 1{,}000{,}000$ random arrangements of the $N = 12$ bivariate observations listed in Table 10.8, there are exactly 1868 δ test statistic values that are equal to or less than the observed value of $\delta = 119.25$.

If all M arrangements of the $N = 12$ bivariate observations listed in Table 10.8 occur with equal chance under the Fisher–Pitman null hypothesis, the Monte Carlo probability value of $\delta = 119.25$ computed on $L = 1{,}000{,}000$ random arrangements of the observed data with $N = 12$ bivariate observations preserved for each arrangement is

$$P(\delta \leq \delta_o | H_0) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{L} = \frac{1868}{1{,}000{,}000} = 0.1868 \times 10^{-2} ,$$

where δ_o denotes the observed value of test statistic δ and L is the number of randomly-selected, equally-likely arrangements of the $N = 12$ bivariate observations listed in Table 10.8.

10.5.2 An Exact Permutation Analysis

While $M = 479{,}001{,}600$ possible arrangements may make an exact permutation analysis impractical, it is not impossible. There are exactly 896,384 δ test statistic values that are equal to or less than the observed value of $\delta = 119.25$. If all M arrangements of the $N = 12$ bivariate observations listed in Table 10.8 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 119.25$ computed on the $M = 479{,}001{,}600$ possible arrangements of the observed data with $N = 12$ bivariate observations preserved for each arrangement is

$$P(\delta \leq \delta_o) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{M} = \frac{896{,}384}{479{,}001{,}600} = 0.1871 \times 10^{-2} ,$$

where δ_o denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 12$ bivariate observations listed

in Table 10.8. Alternatively,

$$P(|r_{xy}| \geq |r_o|) = \frac{\text{number of } |r_{xy}| \text{ values } \geq |r_o|}{M}$$
$$= \frac{896{,}384}{479{,}001{,}600} = 0.1871 \times 10^{-2} ,$$

where $|r_o|$ denotes the observed value of $|r_{xy}|$.

The difference between the exact probability value based on all $M = 479{,}001{,}600$ possible arrangements of the example data listed in Table 10.8 and the Monte Carlo probability value based on $L = 1{,}000{,}000$ random arrangements of the example data is only

$$\Delta_P = 0.001871 - 0.001868 = 0.000003 .$$

To illustrate the accuracy of Monte Carlo permutation methods, Table 10.9 lists 10 independent Monte Carlo analyses of the bivariate data listed in Table 10.8 each initialized with a different seed and each analysis based on $L = 1{,}000{,}000$ random arrangements of the observed data, comparing the Monte Carlo probability values with the exact probability value based on all $M = 479{,}001{,}600$ possible arrangements of the observed data. The exact probability value is $P = 0.001871$, the average of the 10 Monte Carlo probability values listed in Table 10.9 is $P = 0.001868$, and the difference between the average of the 10 Monte Carlo probability values and the exact probability value is

$$\Delta_P = 0.001868 - 0.001871 = 0.000003 ,$$

Table 10.9 Ten independent Monte Carlo runs on the data listed in Table 10.8 based on $L = 1{,}000{,}000$ random arrangements for each run

Run	Seed	Monte Carlo probability	Exact probability	Difference
1	11	0.001912	0.001871	+0.000041
2	13	0.001809	0.001871	−0.000062
3	17	0.001900	0.001871	+0.000029
4	19	0.001896	0.001871	+0.000025
5	23	0.001916	0.001871	+0.000045
6	29	0.001861	0.001871	−0.000010
7	31	0.001809	0.001871	−0.000062
8	37	0.001847	0.001871	−0.000024
9	41	0.001851	0.001871	−0.000020
10	43	0.001883	0.001871	+0.000012

demonstrating the accuracy and efficiency of Monte Carlo permutation statistical methods. Finally, it should be noted that not only are all the differences listed in Table 10.9 very small, but half of the differences are positive and half are negative.

10.6 Example 5: Rank-Score Permutation Analyses

It is not uncommon for researchers to analyze data consisting of rank scores. The correlation coefficients for untied rank-score data most often found in the literature are Spearman's rank-order correlation coefficient given by

$$r_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}\,, \tag{10.2}$$

where for variables x and y, $d_i = x_i - y_i$ for $i = 1, \ldots, N$ bivariate observations, and Kendall's rank-order correlation coefficient given by

$$\tau = \frac{2S}{N(N-1)}\,,$$

where S denotes the number of concordant pairs of rank scores (C) minus the number of discordant pairs (D).[4]

10.6.1 Spearman's Rank-Order Correlation Coefficient

Consider Spearman's rank-order correlation coefficient for N bivariate rank scores under the Neyman–Pearson population model. An example set of data is given in Table 10.10 with $N = 11$ bivariate rank scores.

Following Eq. (10.2) for the data listed in Table 10.10, Spearman's rank-order correlation coefficient is

$$r_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} = 1 - \frac{6(138)}{11(11^2-1)} = +0.3727\,.$$

Under the Neyman–Pearson null hypothesis, H_0: $\rho_{xy} = 0$, Spearman's r_s test statistic is asymptotically distributed as Student's t with $N - 2$ degrees of freedom.

[4]For simplification and clarity the formulæ and examples are limited to untied rank-score data.

Table 10.10 Average weekly spending in dollars on alcohol (x) and tobacco (y) in $N = 11$ Confederate states in 1863

State	Alcohol (x)	Tobacco (y)	Rank x	Rank y	d	d^2
Florida	6.57	2.73	1	11	−10	100
Georgia	6.20	4.48	2	2	0	0
Alabama	6.15	4.51	3	1	+2	4
Mississippi	6.08	3.87	4	4	0	0
Louisiana	5.91	3.54	5	6	−1	1
Arkansas	5.61	3.72	6	5	+1	1
Missouri	5.34	4.21	7	3	+4	16
South Carolina	5.11	2.88	8	10	−2	4
North Carolina	4.87	3.41	9	7	+2	4
Texas	4.49	3.29	10	8	+2	4
Virginia	4.41	3.11	11	9	+2	4
Sum					0	138

For the $N = 11$ bivariate rank scores listed in Table 10.10 with $N - 2 = 11 - 2 = 9$ degrees of freedom,

$$t = r_s\sqrt{\frac{N-2}{1-r_s^2}} = +0.3727\sqrt{\frac{11-2}{1-(+0.3727)^2}} = +1.2050$$

yielding an asymptotic two-tail probability value of $P = 0.2589$, under the assumption of normality.

10.6.2 An Exact Permutation Analysis

For an analysis of the bivariate correlation data listed in Table 10.10 under the Fisher–Pitman permutation model let the differences between the rank scores be squared for correspondence with Spearman's rank-order correlation coefficient. Let $d_i = x_i - y_i$ for $i = 1, \ldots, N$, then the permutation test statistic is given by

$$\delta = \frac{1}{N}\sum_{i=1}^{N} d_i^2 \,. \tag{10.3}$$

Following Eq. (10.3), for the rank-score data listed in Table 10.10 with $N = 11$ bivariate observations the observed value of the permutation test statistic is

$$\delta = \frac{1}{N}\sum_{i=1}^{N} d_i^2 = \frac{138}{11} = 12.5455 \,.$$

Because there are only

$$M = N! = 11! = 39{,}916{,}800$$

possible, equally-likely arrangements in the reference set of all permutations of the alcohol and tobacco data listed in Table 10.10, an exact permutation analysis is feasible. Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the $N = 11$ rank scores listed in Table 10.10 that are equal to or less than the observed value of $\delta = 12.5455$.[5] There are exactly 10,400,726 δ test statistic values that are equal to less than the observed value of $\delta = 12.5455$. If all M arrangements of the $N = 11$ bivariate rank scores listed in Table 10.10 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 12.5455$ computed on the $M = 39{,}916{,}800$ possible arrangements of the observed data with $N = 11$ bivariate observations preserved for each arrangement is

$$P(\delta \leq \delta_o | H_0) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{M} = \frac{10{,}400{,}726}{39{,}916{,}800} = 0.2606\,,$$

where δ_o denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 11$ bivariate rank scores listed in Table 10.10.

10.6.3 The Relationship Between r_s and δ

The functional relationships between test statistics δ and r_s are given by

$$\delta = \frac{(N^2 - 1)(1 - r_s)}{6} \quad \text{and} \quad r_s = 1 - \frac{6\delta}{N^2 - 1}\,. \tag{10.4}$$

Following the first expression given in Eq. (10.4), the observed value of test statistic δ with respect to the observed value of Spearman's r_s is

$$\delta = \frac{(N^2 - 1)(1 - r_s)}{6} = \frac{(11^2 - 1)(1 - 0.3727)}{6} = 12.5455$$

[5] Note that in Eq. (10.3) N is a constant, so only the sum-of-squared differences need be calculated for each arrangement of the observed data.

and, following the second expression in Eq. (10.4), the observed value of Spearman's r_s with respect to the observed value of test statistic δ is

$$r_s = 1 - \frac{6\delta}{N^2 - 1} = 1 - \frac{6(12.5455)}{11^2 - 1} = +0.3727 .$$

Because test statistics δ and r_s are equivalent under the Fisher–Pitman null hypothesis, the exact probability value of Spearman's $r_s = +0.3727$ is identical to the exact probability value of $\delta = 12.5455$; that is,

$$P(\delta \leq \delta_o) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{M} = \frac{10{,}400{,}726}{39{,}916{,}800} = 0.2606$$

and

$$P(|r_s| \geq |r_o|) = \frac{\text{number of } |r_s| \text{ values } \geq |r_o|}{M} = \frac{10{,}400{,}726}{39{,}916{,}800} = 0.2606 ,$$

where δ_o and r_o denote the observed values of δ and r_s, respectively, and M is the number of possible, equally-likely arrangements of the $N = 11$ bivariate rank scores listed in Table 10.10.

The exact expected value of the $M = 39{,}916{,}800$ δ test statistic values under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i = \frac{798{,}336{,}000}{39{,}916{,}800} = 20.00 .$$

Alternatively, the exact expected value of test statistic δ is

$$\mu_\delta = \frac{N^2 - 1}{6} = \frac{11^2 - 1}{6} = 20.00 .$$

Then the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{12.5455}{20.00} = +0.3727 ,$$

indicating approximately 37% agreement between the x and y rank-score values above what is expected by chance.

When the N rank-score values in variable y are a simple permutation of the rank-score values in variable x it can easily be shown that Mielke and Berry's $\Re$ measure of effect size and Spearman's r_s rank-order correlation coefficient are equivalent under the Neyman–Pearson population model with squared Euclidean scaling; that

is, $\Re = +0.3727$ and $r_s = +0.3727$. Specifically, given

$$\delta = \frac{(N^2-1)(1-r_s)}{6} \quad \text{and} \quad \mu_\delta = \frac{N^2-1}{6} ,$$

then

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{(N^2-1)(1-r_s)}{6} \times \frac{6}{N^2-1} = 1 - (1-r_s) = r_s .$$

10.6.4 Kendall's Rank-Order Correlation Coefficient

A popular alternative to Spearman's r_s rank-order correlation coefficient is Kendall's τ rank-order correlation coefficient given by

$$\tau = \frac{S}{\binom{N}{2}} = \frac{2S}{N(N-1)} ,$$

where $S = C - D$, C denotes the number of concordant pairs of the observed data, and D denotes the number of discordant pairs of the observed data. To illustrate the difference between concordant and discordant pairs, consider the example data with $N = 4$ bivariate rank scores listed in Table 10.11. There are

$$M = \binom{N}{2} = \binom{4}{2} = \frac{4(4-1)}{2} = 6$$

possible, equally-likely arrangements in the reference set of all permutations of the example data listed in Table 10.11 to be considered. The first (x,y) pair is $x_1 = 1$ and $x_2 = 2$, and $y_1 = 2$ and $y_2 = 3$. Since $x_1 = 1$ is less than $x_2 = 2$ and $y_1 = 2$ is less than $y_2 = 3$, the first (x,y) pair is considered to be *concordant* as the values of variables x and y are in the same order for the pair.

The next (x,y) pair is $x_1 = 1$ and $x_3 = 3$, and $y_1 = 2$ and $y_3 = 1$. Since $x_1 = 1$ is less than $x_3 = 3$ but $y_1 = 2$ is greater than $y_3 = 1$, the second (x,y)

Table 10.11 Example rank-score data on $N = 4$ bivariate observations

	Variable	
Object	x	y
1	1	2
2	2	3
3	3	1
4	4	4

Table 10.12 Calculation of concordant (C) and discordant (D) pairs for the example rank-score data listed in Table 10.11

	Variable			
Pair	x	y	C	D
1	$1 < 2$	$2 < 3$	1	
2	$1 < 3$	$2 > 1$		1
3	$1 < 4$	$2 < 4$	1	
4	$2 < 3$	$3 > 1$		1
5	$2 < 4$	$3 < 4$	1	
6	$3 < 4$	$1 < 4$	1	
Sum			4	2

pair is considered to be *discordant* as the values of variables x and y are not in the same order for the pair. Table 10.12 illustrates the calculation of the six concordant (C) and discordant (D) pairs for the rank-score data listed in Table 10.11. For the six (x,y) pairs listed in Table 10.12, the number of concordant pairs is $C = 4$, the number of discordant pairs is $D = 2$, and Kendall's $S = C - D = 4 - 2 = +2$. When M becomes large the calculations can become cumbersome. Table 10.13 illustrates the calculation of Kendall's S for the rank-score data listed in Table 10.12 on p. 388 with $N = 11$ bivariate pairs.

The process illustrated in Table 10.13 is straightforward. Determine the value for Kendall's S by arranging the values of variable x in their natural order and arranging the values of variable y corresponding to the values of variable x, as in Table 10.13. Starting with the first value of variable y on the left (11), count the number of rank scores to the right of 11 that are smaller than 11 and score each as (-1); these (-1) values represent the disagreements in order. For the calculations listed in Table 10.13 there are 10 values that are smaller than 11. Next count the number of rank scores to the right of 11 that are larger than 11 and score each as $(+1)$; these $(+1)$ values

Table 10.13 Example calculations for determining the value of Kendall's S test statistic

x	1	2	3	4	5	6	7	8	9	10	11	
y	11	2	1	4	6	5	3	10	7	8	9	Sum
		−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−10
			−1	+1	+1	+1	+1	+1	+1	+1	+1	+7
				+1	+1	+1	+1	+1	+1	+1	+1	+8
					+1	+1	−1	+1	+1	+1	+1	+5
						−1	−1	+1	+1	+1	+1	+2
							−1	+1	+1	+1	+1	+3
								+1	+1	+1	+1	+4
									−1	−1	−1	−3
										+1	+1	+2
											+1	+1
												0
Sum												+19

represent the agreements in order. In this case there are no values larger than 11. Sum the 10 (−1) and zero (+1) values and place the sum at the end of the first row.

The next y value is 2. Count the number of rank scores to the right of 2 that are smaller than 2 and score each as (−1); there is only one value (1) that is smaller than 2. Next count the number of rank scores to the right of 2 that are larger than 2 and score each as (+1); there are eight values that are larger than 2. Sum the one (−1) and eight (+1) values and place the sum at the end of the second row. Continue the procedure for all ranks in variable y, summing the results. The final sum is the value for Kendall's S. Alternatively, there are 37 (+1) values in Table 10.13; these are the concordant pairs (C). There are 18 (−1) values in Table 10.13; these are the discordant pairs (D). Then, $S = C - D = 37 - 18 = +19$.

For the rank-score data listed in Table 10.10 on p. 384 with $N = 11$ untied rank scores, Kendall's rank-order correlation coefficient is

$$\tau = \frac{2S}{N(N-1)} = \frac{2(+19)}{11(11-1)} = +0.3455 \, .$$

In testing the significance of the association between paired ranks it is more convenient to apply a test directly to S rather than τ as the number of pairs, $N(N-1)/2$, is a constant. Kendall's S is asymptotically distributed $N(0, 1)$ with mean of zero and variance given by

$$\sigma_S^2 = \frac{N(N-1)(2N+5)}{18}$$

as $N \to \infty$. Since the normal distribution is an approximation to the discrete sampling distribution of S, a correction for continuity should be applied. For the rank-score data listed in Table 10.10 on p. 384, the normal deviate with continuity correction applied is

$$z = \frac{|S| - 1}{\left[N(N-1)(2N+5)/18\right]^{1/2}}$$

$$= \frac{19 - 1}{\left\{11(11-1)[(2)(11)+5]/18\right\}^{1/2}} = +1.4013 \, ,$$

yielding an asymptotic two-tail probability value of $P = 0.1611$, under the assumption of normality.

10.6.5 An Exact Permutation Analysis

Consider an analysis of the correlation data listed in Table 10.10 on p. 384 with $N = 11$ bivariate observations under the Fisher–Pitman permutation model. There

are

$$M = N! = 11! = 39{,}916{,}800$$

possible, equally-likely arrangements in the reference set of all permutations of the example data listed in Table 10.10, making an exact permutation analysis feasible. Under the Fisher–Pitman permutation model, the exact probability of an observed value of Kendall's S is the proportion of S test statistic values computed on all possible, equally-likely arrangements of the $N = 11$ bivariate rank scores listed in Table 10.10 that are equal to or greater than the observed value of $S = +19$. There are exactly 6,436,200 S test statistic values that are equal to or greater than the observed value of $S = +19$. If all M arrangements of the $N = 11$ bivariate rank scores listed in Table 10.10 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $S = 19$ computed on the $M = 39{,}916{,}800$ possible arrangements of the observed data with $N = 11$ bivariate observations preserved for each arrangement is

$$P(|S| \geq |S_o|) = \frac{\text{number of } |S| \text{ values } \geq |S_o|}{M} = \frac{6{,}436{,}200}{39{,}916{,}800} = 0.1612\,,$$

where S_o denotes the observed value of Kendall's S and M is the number of possible, equally-likely arrangements of the $N = 11$ bivariate observations listed in Table 10.10.

10.6.6 Spearman's Footrule Correlation Coefficient

While Charles Spearman is most often remembered for his contributions to factor analysis and his development of the rank-order correlation coefficient given by

$$r_s = 1 - \frac{6\sum_{i=1}^{2} d_i^2}{N(N^2 - 1)}\,,$$

which was discussed in Sect. 10.6.1, Spearman also developed a lesser-known correlation coefficient that he called the "footrule" given by

$$\mathcal{R} = 1 - \frac{3\sum_{i=1}^{N} |x_i - y_i|}{N^2 - 1}\,,$$

where x_i and y_i denote the ith observed rank-score values for $i = 1, \ldots, N$ and N is the number of bivariate rank scores.

Table 10.14 Example bivariate rank-score correlation data with $N = 8$ pairs of data

Pair	x	y	$x - y$	$\lvert x - y \rvert$
1	8	7	+1	1
2	6	6	0	0
3	2	4	−2	2
4	4	2	+2	2
5	7	8	−1	1
6	5	5	0	0
7	1	3	−2	2
8	3	1	+2	2
Sum				10

To illustrate Spearman's footrule measure of correlation, consider the example data listed in Table 10.14 with $N = 8$ bivariate untied rank-score observations. For the $N = 8$ bivariate rank-score observations listed in Table 10.14, Spearman's footrule is

$$\mathcal{R} = 1 - \frac{3\sum_{i=1}^{N} |x_i - y_i|}{N^2 - 1} = 1 - \frac{3(10)}{8^2 - 1} = +0.5238 \, .$$

For comparison, Spearman's rank-order correlation coefficient calculated on the rank-score data listed in Table 10.14 is $r_s = +0.7857$ and Kendall's rank-order correlation coefficient is $\tau = +0.6429$.

Since there are only

$$M = N! = 8! = 40{,}320$$

possible, equally-likely arrangements in the reference set of all permutations of the observed x and y rank scores listed in Table 10.14, an exact permutation analysis is feasible. Under the Fisher–Pitman permutation model, the exact probability of an observed $\mathcal{R}$ is the proportion of $\mathcal{R}$ test statistic values computed on all possible, equally-likely arrangements of the $N = 8$ bivariate rank scores listed in Table 10.14 that are equal to or greater than the observed value of $\mathcal{R} = +0.5238$. There are exactly 1248 $\mathcal{R}$ test statistic values that are equal to or greater than the observed value of $\mathcal{R} = +0.5238$. If all M arrangements of the $N = 8$ rank scores listed in Table 10.14 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\mathcal{R} = +0.5238$ computed on the $M = 40{,}320$ possible arrangements of the observed data with $N = 8$ bivariate rank scores preserved for each arrangement is

$$P(\mathcal{R} \geq \mathcal{R}_o | H_0) = \frac{\text{number of } \mathcal{R} \text{ values } \geq \mathcal{R}_o}{M} = \frac{1248}{40{,}320} = 0.0310 \, ,$$

where $\mathcal{R}_o$ denotes the observed value of Spearman's $\mathcal{R}$ and M is the number of possible, equally-likely arrangements of the $N = 8$ bivariate rank scores listed in Table 10.14.

10.6.7 The Relationship Between Statistics $\mathcal{R}$ and $\Re$

It can easily be demonstrated that Spearman's $\mathcal{R}$ footrule measure and Mielke and Berry's $\Re$ measure of effect size are equivalent measures under the Fisher–Pitman permutation model with ordinary Euclidean scaling. Let

$$\delta = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{10.5}$$

denote an average distance function based on all possible paired absolute differences among values of the two rankings and let

$$\mu_\delta = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - y_j| \tag{10.6}$$

denote the expected value of test statistic δ. Then Spearman's footrule measure is given by

$$\mathcal{R} = 1 - \frac{\delta}{\mu_\delta}, \tag{10.7}$$

which is also the equation for Mielke and Berry's $\Re$ measure of effect size.

The calculation of test statistics δ, μ_δ, and $\Re$ can be illustrated and compared with Spearman's $\mathcal{R}$ footrule measure using an example set of data. Consider the small set of rank-score data listed in Table 10.15 with $N = 5$ bivariate observations. Table 10.16 illustrates the calculation of Spearman's footrule measure for the rank-score data listed in Table 10.15. Given the calculations listed in Table 10.16, the

Table 10.15 Bivariate rank scores assigned to $N = 5$ objects

Object	x	y
1	5	4
2	2	1
3	1	2
4	3	3
5	4	5

Table 10.16 Detailed calculations for Spearman's footrule measure with $N = 5$ bivariate observations

Pair	i	x_i	y_i	$x_i - y_i$	$\lvert x_i - y_i \rvert$
1	1	5	4	−1	1
2	2	2	1	+1	1
3	3	1	2	−1	1
4	4	3	3	0	0
5	5	4	5	−1	1

Table 10.17 Calculation of $\lvert x_i - y_i \rvert$ for $i = 1, \ldots, N$ for δ

Pair	i	x_i	y_i	$\lvert x_i - y_i \rvert$
1	1	5	4	$\lvert 5 - 4 \rvert = 1$
2	2	2	1	$\lvert 2 - 1 \rvert = 1$
3	3	1	2	$\lvert 1 - 2 \rvert = 1$
4	4	3	3	$\lvert 3 - 3 \rvert = 0$
5	5	4	5	$\lvert 4 - 5 \rvert = 1$

observed value of Spearman's footrule measure is

$$\mathcal{R} = \frac{3\sum_{i=1}^{N} \lvert x_i - y_i \rvert}{N^2 - 1} = \frac{3(1+1+1+0+1)}{5^2 - 1} = +0.50 \, .$$

Table 10.17 illustrates the calculation of δ for the rank-score data listed in Table 10.15. Given the calculations listed in Table 10.17, the observed value of test statistic δ is

$$\delta = \frac{1}{N} \sum_{i=1}^{N} \lvert x_i - y_i \rvert = \frac{1+1+1+0+1}{5} = 0.80 \, .$$

Table 10.18 illustrates the calculation of μ_δ for the rank-score data listed in Table 10.15. Given the calculations listed in Table 10.18, the exact expected value of the N^2 δ test statistic values under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lvert x_i - y_j \rvert = \frac{1+0+2+\cdots+2+0+1}{5^2} = 1.60 \, .$$

Then the chance-corrected measure of agreement is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{0.80}{1.60} = +0.50 \, ,$$

Table 10.18 Calculation of $|x_i - y_j|$ for $i, j = 1, \ldots, N$ for μ_δ

Pair	i	j	$\lvert x_i - y_j \rvert$	Pair	i	j	$\lvert x_i - y_j \rvert$
1	1	2	$\lvert 1-2 \rvert = 1$	14	3	5	$\lvert 3-5 \rvert = 2$
2	1	1	$\lvert 1-1 \rvert = 0$	15	3	4	$\lvert 3-4 \rvert = 1$
3	1	3	$\lvert 1-3 \rvert = 2$	16	4	2	$\lvert 4-2 \rvert = 2$
4	1	5	$\lvert 1-5 \rvert = 4$	17	4	1	$\lvert 4-1 \rvert = 3$
5	1	4	$\lvert 1-4 \rvert = 3$	18	4	3	$\lvert 4-3 \rvert = 1$
6	2	2	$\lvert 2-2 \rvert = 0$	19	4	5	$\lvert 4-5 \rvert = 1$
7	2	1	$\lvert 2-1 \rvert = 1$	20	4	4	$\lvert 4-4 \rvert = 0$
8	2	3	$\lvert 2-3 \rvert = 1$	21	5	2	$\lvert 5-2 \rvert = 3$
9	2	5	$\lvert 2-5 \rvert = 3$	22	5	1	$\lvert 5-1 \rvert = 4$
10	2	4	$\lvert 2-4 \rvert = 2$	23	5	3	$\lvert 5-3 \rvert = 2$
11	3	2	$\lvert 3-2 \rvert = 1$	24	5	5	$\lvert 5-5 \rvert = 0$
12	3	1	$\lvert 3-1 \rvert = 2$	25	5	4	$\lvert 5-4 \rvert = 1$
13	3	3	$\lvert 2-5 \rvert = 0$				

indicating 50% agreement above that expected by chance. Thus, the equivalence between

$$\mathcal{R} = 1 - \frac{3\sum_{i=1}^{N} |x_i - y_i|}{N^2 - 1} \quad \text{and} \quad \Re = 1 - \frac{\delta}{\mu_\delta}$$

is demonstrated.

10.6.8 A More Rigorous Proof

In this section a proof is offered that mathematically establishes the equivalence of Spearman's footrule measure and Mielke and Berry's chance-corrected measure of effect size. Consider the expected value of test statistic δ as defined in Eq. (10.6) and given by

$$\mu_\delta = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| x_i - y_j \right| .$$

Then,

$$\begin{aligned} \mu_\delta &= \frac{2}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (j - i) \\ &= \frac{1}{N^2} \sum_{i=1}^{N-1} \left[N(N+1) + i^2 - i(2N+1) \right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{N(N-1)}{6N^2}\big[6(N+1) + (2N-1) - 3(2N+1)\big] \\
&= \frac{N-1}{6N}\big[2(N+1)\big] \\
&= \frac{N^2-1}{3N}
\end{aligned}
$$

The chance-corrected measure of effect size defined in Eq. (10.7) on p. 392 is

$$\Re = 1 - \frac{\delta}{\mu_\delta} \, .$$

Therefore,

$$\delta = \mu_\delta(1 - \Re) \, .$$

Given the permutation test statistic defined in Eq. (10.5) on p. 392; that is,

$$\delta = \frac{1}{N}\sum_{i=1}^{N} \big|x_i - y_i\big| \, ,$$

and substituting δ into Spearman's footrule measure

$$\mathcal{R} = 1 - \frac{3\sum_{i=1}^{N} \big|x_i - y_i\big|}{N^2 - 1}$$

yields

$$\mathcal{R} = 1 - \frac{3N\delta}{N^2 - 1}$$

and substituting $\mu_\delta(1 - \Re)$ for δ yields

$$\mathcal{R} = 1 - \frac{3N\mu_\delta(1 - \Re)}{N^2 - 1} \, .$$

Finally, substituting $(N^2 - 1)/3N$ for μ_δ yields

$$\mathcal{R} = 1 - \frac{3N\left(\dfrac{N^2-1}{3N}\right)(1 - \Re)}{N^2 - 1} = 1 - (1 - \Re) = \Re \, .$$

10.7 Example 6: Multivariate Permutation Analyses

Many introductory textbooks in statistics include a brief introduction to multiple correlation, usually limiting the discussion to two predictors for simplicity. The OLS multiple regression equation is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \ ,$$

where $\hat{y}$ denotes the predicted value of the criterion variable, $x_1, \ x_2, \ \ldots, \ x_p$ denote p predictor variables, $\hat{\beta}_1, \ \hat{\beta}_2, \ \ldots, \ \hat{\beta}_p$ denote the OLS regression weights for each of the p predictor variables, and $\hat{\beta}_0$ is the estimate of the population intercept. The assumptions underlying OLS multiple regression are (1) the observations are independent, (2) a linear relationship exists between the criterion variable and the predictor variables, (3) multivariate normality, (4) no multicollinearity among the variables, and (5) the variances of the error terms are similar across the values of the p predictor variables; that is, homogeneity.

10.7.1 A Conventional OLS Multivariate Analysis

To illustrate multiple correlation analyses with OLS and LAD regression, consider the example data listed in Table 10.19 with $p = 2$ predictors where variable y is Hours of Housework done by husbands per week, variable x_1 is Number of Children in the family, and variable x_2 is husband's Years of Education for $N = 12$ families. For the multivariate data listed in Table 10.19, the unstandardized OLS regression coefficients are

$$\hat{\beta}_0 = +2.5260 \ , \quad \hat{\beta}_1 = +0.6356 \ , \quad \text{and} \quad \hat{\beta}_2 = -0.0649 \ ,$$

Table 10.19 Example multivariate correlation data on $N = 12$ families with $p = 2$ predictors

Family	x_1	x_2	y
A	1	12	1
B	1	14	2
C	1	16	3
D	1	16	5
E	2	18	3
F	2	16	1
G	3	12	5
H	3	12	0
I	4	10	6
J	4	12	3
K	5	10	7
L	5	16	4

Table 10.20 Observed x and y values with associated predicted values ($\hat{y}$), residual errors ($\hat{e}$), and squared residual errors ($\hat{e}^2$) for the multivariate correlation data listed in Table 10.19

Family	x_1	x_2	y	$\hat{y}$	$\hat{e}$	$\hat{e}^2$
A	1	12	1	2.3823	−1.3823	1.9108
B	1	14	2	2.2525	−0.2525	0.0638
C	1	16	3	2.1226	+0.8774	0.7698
D	1	16	5	2.1226	+2.8774	8.2795
E	2	18	3	2.6283	+0.3717	0.1382
F	2	16	1	2.7581	−1.7582	3.0911
G	3	12	5	3.6534	+1.3466	1.8132
H	3	12	0	3.6534	−3.6534	13.3477
I	4	10	6	4.4189	+1.5811	2.5000
J	4	12	3	4.2889	−1.2890	1.6615
K	5	10	7	5.0544	+1.9456	3.7853
L	6	16	4	4.6648	−0.6648	0.4420
Sum	32	164	40	40.0000	0.0000	37.8028

and the observed squared OLS multiple correlation coefficient is $R^2 = 0.2539$. Table 10.20 lists the $N = 12$ observed values for variables x and y, the predicted y values ($\hat{y}$), the residual errors ($\hat{e}$), and the squared residual errors ($\hat{e}^2$).

The summary statistics given in Table 10.20 suggest an alternative method to determine the value of the multiple correlation coefficient. Define

$$R^2 = r^2_{y\hat{y}} = \frac{\left[N\sum_{i=1}^{N} y\hat{y} - \left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} \hat{y}\right)\right]^2}{\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]\left[N\sum_{i=1}^{N} \hat{y}_i^2 - \left(\sum_{i=1}^{N} \hat{y}_i\right)^2\right]} . \tag{10.8}$$

For the multivariate data listed in Table 10.19, $N = 12$,

$$\sum_{i=1}^{N} y_i = 40.00 , \quad \sum_{i=1}^{N} y_i^2 = 184.00 \quad \sum_{i=1}^{N} \hat{y}_i = 40.00 , \quad \sum_{i=1}^{N} \hat{y}_i^2 = 146.1984 ,$$

and

$$\sum_{i=1}^{N} y\hat{y} = 146.1984 .$$

Then following Eq. (10.8),

$$R^2 = r^2_{y\hat{y}} = \frac{\left[12(146.1984) - (40)(40)\right]^2}{\left[12(184.00) - (40)^2\right]\left[12(146.1984) - (40)^2\right]} = 0.2539 .$$

If, under the Neyman–Pearson population model the null hypothesis posits the population correlation is zero; that is, H_0: $R_{y \cdot x_1, x_2} = 0$, the conventional OLS test of significance is given by

$$F = \frac{R^2(N - p - 1)}{p(1 - R^2)} ,$$

which is asymptotically distributed as Snedecor's F with $\nu_1 = p$ and $\nu_2 = N - p - 1$ degrees of freedom. For the multivariate data listed in Table 10.19,

$$F = \frac{R^2(N - p - 1)}{p(1 - R^2)} = \frac{0.2539(12 - 2 - 1)}{2(1 - 0.2539)} = 1.5313$$

and with $\nu_1 = p = 2$ and $\nu_2 = N - p - 1 = 12 - 2 - 1 = 9$ degrees of freedom, the asymptotic probability value of $F = 1.5313$ is $P = 0.2677$, under the assumptions of linearity, normality, and homogeneity.

10.7.2 A Monte Carlo Permutation Analysis

Because there are

$$M = N! = 12! = 479{,}001{,}600$$

possible, equally-likely arrangements in the reference set of all permutations of the family data listed in Table 10.19, a Monte Carlo permutation analysis is most appropriate. Under the Fisher–Pitman permutation model, the Monte Carlo probability of an observed R^2 is the proportion of R^2 test statistic values computed on the randomly-selected, equally-likely arrangements of the $N = 12$ multivariate observations listed in Table 10.19 that are equal to or greater than the observed value of $R^2 = 0.2539$. Based on $L = 1{,}000{,}000$ randomly-selected arrangements of the $N = 12$ multivariate observations listed in Table 10.19, there are exactly 268,026 R^2 test statistic values that are equal to greater than the observed value of $R^2 = 0.2539$.

If all M arrangements of the $N = 12$ multivariate observations listed in Table 10.19 occur with equal chance under the Fisher–Pitman null hypothesis, the Monte Carlo probability value of $R^2 = 0.2539$ computed on $L = 1{,}000{,}000$ randomly-selected arrangements of the observed data with $N = 12$ multivariate observations preserved for each arrangement is

$$P\left(R^2 \geq R^2_{\mathrm{o}} | H_0\right) = \frac{\text{number of } R^2 \text{ values } \geq R^2_{\mathrm{o}}}{L} = \frac{268{,}026}{1{,}000{,}000} = 0.2680 ,$$

where R^2_o denotes the observed value of R^2 and L is the number of randomly-selected, equally-likely arrangements of the multivariate observations listed in Table 10.19.

10.7.3 An Exact Permutation Analysis

While $M = 479{,}001{,}600$ possible arrangements may make an exact permutation analysis impractical, it is not impossible. There are exactly 128,420,329 R^2 test statistic values that are equal to or greater than the observed value of $R^2 = 0.2539$. If all M arrangements of the $N = 12$ multivariate observations listed in Table 10.19 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $R^2 = 0.2539$ computed on the $M = 479{,}001{,}600$ possible arrangements of the observed data with $N = 12$ multivariate observations preserved for each arrangement is

$$P\left(R^2 \geq R^2_o | H_0\right) = \frac{\text{number of } R^2 \text{ values } \geq R^2_o}{M} = \frac{128{,}420{,}329}{479{,}001{,}600} = 0.2681 \,,$$

where R^2_o denotes the observed value of R^2 and M is the number of possible, equally-likely arrangements of the multivariate observations listed in Table 10.19.

10.7.4 A LAD Multivariate Regression Analysis

Now consider a LAD regression analysis of the multivariate data listed in Table 10.19 on p. 396. Table 10.21 lists the $N = 12$ observed values for variables x_1, x_2, and y, the predicted y values ($\tilde{y}$), the residual errors ($\tilde{e}$), and the absolute residual errors ($|\tilde{e}|$).

For the family data listed in Table 10.19, the LAD regression coefficients are

$$\tilde{\beta}_0 = +4.7500 \,, \quad \tilde{\beta}_1 = +0.2500 \,, \quad \text{and} \quad \tilde{\beta}_2 = -0.1250 \,,$$

the observed permutation test statistic is

$$\delta = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \tilde{y}_i \right| = \frac{1}{N} \sum_{i=1}^{N} |\tilde{e}| = \frac{18}{12} = 1.50 \,,$$

Table 10.21 Observed x and y values with associated predicted values ($\tilde{y}$), residual errors ($\tilde{e}$), and absolute errors ($|\tilde{e}|$) for the multivariate correlation data listed in Table 10.19

Family	x_1	x_2	y	$\tilde{y}$	$\tilde{e}$	$\lvert\tilde{e}\rvert$
A	1	12	1	3.5000	−2.5000	2.5000
B	1	14	2	3.2500	−1.2500	1.2500
C	1	16	3	3.0000	0.0000	0.0000
D	1	16	5	3.0000	+2.0000	2.0000
E	2	18	3	3.0000	0.0000	0.0000
F	2	16	1	3.2500	−2.2500	2.2500
G	3	12	5	4.0000	+1.0000	1.0000
H	3	12	0	4.0000	−4.0000	4.0000
I	4	10	6	4.5000	+1.5000	1.5000
J	4	12	3	4.2500	−1.2500	1.2500
K	5	10	7	4.7500	+2.2500	2.2500
L	6	16	4	4.0000	0.0000	0.0000
Sum	32	164	40	44.5000	−4.5000	18.0000

the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis is

$$\begin{aligned}\mu_\delta &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|y_i - \tilde{y}_j\right| \\ &= \frac{|1-3.50|+|1-3.25|+|1-3.00|+\cdots+|4-4.75|+|4-4.00|}{12^2} \\ &= \frac{260}{144} = 1.8056\,,\end{aligned}$$

and the observed LAD measure of agreement between the y and $\tilde{y}$ values is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{1.5000}{1.8056} = +0.1692\,,$$

indicating approximately 17% agreement between the observed and predicted y values.

There are

$$M = N! = 12! = 479{,}001{,}600$$

possible, equally-likely arrangements in the reference set of all permutations of the family data listed in Table 10.19, making an exact permutation analysis impractical. Under the Fisher–Pitman permutation model, the Monte Carlo probability of an observed $\Re$ is the proportion of $\Re$ test statistic values computed on the randomly-selected, equally-likely arrangements of the $N = 12$ multivariate observations

listed in Table 10.19 that are equal to or greater than the observed value of $\Re = +0.1692$. Based on $L = 1{,}000{,}000$ randomly-selected arrangements of the $N = 12$ multivariate observations listed in Table 10.19, there are exactly 37,824 $\Re$ test statistic values that are equal to greater than the observed value of $\Re = +0.1692$.

If all M arrangements of the $N = 12$ multivariate observations listed in Table 10.19 occur with equal chance under the Fisher–Pitman null hypothesis, the Monte Carlo probability value of $\Re = +0.1692$ computed on $L = 1{,}000{,}000$ randomly-selected arrangements of the observed data with $N = 12$ multivariate observations preserved for each arrangement is

$$P\left(\Re \geq \Re_{\mathrm{o}} | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_{\mathrm{o}}}{L} = \frac{37{,}824}{1{,}000{,}000} = 0.0378\,,$$

where $\Re_{\mathrm{o}}$ denotes the observed value of $\Re$ and L is the number of randomly-selected, equally-likely arrangements of the $N = 12$ multivariate observations listed in Table 10.19.

10.7.5 An Exact Permutation Analysis

Now consider an exact permutation analysis of the $M = 479{,}001{,}600$ arrangements of the family data listed in Table 10.19. If all M arrangements of the $N = 12$ multivariate observations listed in Table 10.19 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\Re = +0.1692$ computed on the $M = 479{,}001{,}600$ possible arrangements of the observed data with $N = 12$ multivariate observations preserved for each arrangement is

$$P\left(\Re \geq \Re_{\mathrm{o}} | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_{\mathrm{o}}}{M} = \frac{18{,}117{,}645}{479{,}001{,}600} = 0.0378\,,$$

where $\Re_{\mathrm{o}}$ denotes the observed value of $\Re$ and M is the number of possible, equally-likely arrangements of the $N = 12$ multivariate observations listed in Table 10.19.

10.7.6 Analyses with an Extreme Value

Suppose that the husband in Family "L" in Table 10.19 on p. 396 was a stay-at-home house-husband and instead of contributing just 4 h of housework per week, he actually contributed 40 h, as in Table 10.22.

Table 10.22 Example multivariate correlation data on $N = 12$ families with $p = 2$ predictors, where the husband in family L contributed 40 h of housework per week

Family	x_1	x_2	y
A	1	12	1
B	1	14	2
C	1	16	3
D	1	16	5
E	2	18	3
F	2	16	1
G	3	12	5
H	3	12	0
I	4	10	6
J	4	12	3
K	5	10	7
L	5	16	40

10.7.7 An Ordinary Least Squares (OLS) Analysis

For the multivariate data listed in Table 10.22, the unstandardized OLS regression coefficients are

$$\hat{\beta}_0 = -41.6558\,, \quad \hat{\beta}_1 = +5.7492\,, \quad \text{and} \quad \hat{\beta}_2 = +2.3896\,,$$

and the observed squared OLS multiple correlation coefficient is $R^2 = 0.5786$.

There are

$$M = N! = 12! = 479{,}001{,}600$$

possible, equally-likely arrangements in the reference set of all permutations of the family data listed in Table 10.22, making an exact permutation analysis impractical. Under the Fisher–Pitman permutation model, the Monte Carlo probability of an observed R^2 is the proportion of R^2 test statistic values computed on the randomly-selected, equally-likely arrangements of the observed data that are equal to or greater than the observed value of $R^2 = 0.5786$. Based on $L = 1{,}000{,}000$ randomly-selected arrangements of the $N = 12$ multivariate observations listed in Table 10.22, there are exactly 15,215 R^2 test statistic values that are equal to greater than the observed value of $R^2 = 0.5786$.

If all M arrangements of the $N = 12$ multivariate observations listed in Table 10.22 occur with equal chance under the Fisher–Pitman null hypothesis, the Monte Carlo probability value of $R^2 = 0.5786$ computed on $L = 1{,}000{,}000$ random arrangements of the observed data with $N = 12$ multivariate observations preserved for each arrangement is

$$P\left(R^2 \geq R^2_{\mathrm{o}} | H_0\right) = \frac{\text{number of } R^2 \text{ values } \geq R^2_{\mathrm{o}}}{L} = \frac{15{,}215}{1{,}000{,}000} = 0.0152\,,$$

where R^2_o denotes the observed value of R^2 and L is the number of randomly-selected, equally-likely arrangements of the $N = 12$ multivariate observations listed in Table 10.22.

Although an exact permutation analysis of $M = 479{,}001{,}600$ arrangements of the family data listed in Table 10.22 may be impractical, it is not impossible. If all M arrangements of the $N = 12$ multivariate observations listed in Table 10.22 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $R^2 = 0.5786$ computed on the $M = 479{,}001{,}600$ possible arrangements of the observed data with $N = 12$ multivariate observations preserved for each arrangement is

$$P\left(R^2 \geq R^2_o | H_0\right) = \frac{\text{number of } R^2 \text{ values } \geq R^2_o}{M} = \frac{7{,}328{,}725}{479{,}001{,}600} = 0.0153 \, ,$$

where R^2_o denotes the observed value of R^2 and M is the number of possible, equally-likely arrangements of the $N = 12$ multivariate observations listed in Table 10.22.

For comparison,

$$F = \frac{R^2(N - p - 1)}{p(1 - R^2)} = \frac{0.5786(12 - 2 - 1)}{2(1 - 0.5786)} = 6.1785 \, ,$$

where F is asymptotically distributed as Snedecor's F with $\nu_1 = p$ and $\nu_2 = N - p - 1$ degrees of freedom. With $\nu_1 = p = 2$ and $\nu_2 = N - p - 1 = 12 - 2 - 1 = 9$ degrees of freedom, the asymptotic probability value of $F = 6.1785$ is $P = 0.0205$, under the assumptions of linearity, normality, and homogeneity.

10.7.8 A Least Absolute Deviation (LAD) Analysis

For the multivariate family data listed in Table 10.22 on p. 402, the LAD regression coefficients are

$$\tilde{\beta}_0 = -6.75 \, , \quad \tilde{\beta}_1 = +1.75 \, , \quad \tilde{\beta}_2 = +0.50 \, ,$$

the observed permutation test statistic is $\delta = 3.9583$, the exact expected value of δ under the Fisher–Pitman null hypothesis is $\mu_\delta = 5.4687$, and the LAD chance-corrected measure of agreement between the observed y values and the predicted $\tilde{y}$ values is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{3.9583}{5.4687} = +0.2762 \, ,$$

indicating approximately 28% agreement between the observed and predicted y values.

There are

$$M = N! = 12! = 479{,}001{,}600$$

possible, equally-likely arrangements in the reference set of all permutations of the family data listed in Table 10.22, making an exact permutation analysis impractical. Under the Fisher–Pitman permutation model, the Monte Carlo probability of an observed $\Re$ is the proportion of $\Re$ test statistic values computed on the randomly-selected, equally-likely arrangements of the observed data that are equal to or greater than the observed value of $\Re = +0.2762$. Based on $L = 1{,}000{,}000$ randomly-selected arrangements of the $N = 12$ multivariate observations listed in Table 10.22, there are exactly 3409 $\Re$ test statistic values that are equal to greater than the observed value of $\Re = +0.2762$.

If all M arrangements of the $N = 12$ multivariate observations listed in Table 10.22 occur with equal chance under the Fisher–Pitman null hypothesis, the Monte Carlo probability value of $\Re = +0.2762$ computed on $L = 1{,}000{,}000$ randomly-selected arrangements of the observed data with $N = 12$ multivariate observations preserved for each arrangement is

$$P\left(\Re \geq \Re_{\mathrm{o}} | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_{\mathrm{o}}}{L} = \frac{3409}{1{,}000{,}000} = 0.3409 \times 10^{-2},$$

where $\Re_{\mathrm{o}}$ denotes the observed value of $\Re$ and L is the number of randomly-selected, equally-likely arrangements of the $N = 12$ multivariate observations listed in Table 10.22.

For comparison, consider an exact permutation analysis of the $M = 479{,}001{,}600$ arrangements of the observed data. If all M arrangements of the $N = 12$ multivariate observations listed in Table 10.22 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\Re = +0.2762$ computed on the $M = 479{,}001{,}600$ possible arrangements of the observed data with $N = 12$ multivariate observations preserved for each arrangement is

$$P\left(\Re \geq \Re_{\mathrm{o}} | H_0\right) = \frac{\text{number of } \Re \text{ values } \geq \Re_{\mathrm{o}}}{M} = \frac{163{,}234{,}242}{479{,}001{,}600} = 0.3408 \times 10^{-2},$$

where $\Re_{\mathrm{o}}$ denotes the observed value of $\Re$ and M is the number of possible, equally-likely arrangements of the $N = 12$ multivariate observations listed in Table 10.22.

The results of the comparison of the OLS and LAD regression analyses with $y_{12} = 4$ and $y_{12} = 40$ h of housework by the husband in family "L" are summarized

Table 10.23 Comparison of OLS and LAD analyses for the data given in Table 10.19 with 4 h of housework for the husband in family *L* and the data given in Table 10.22 with 40 h of housework for the husband in family *L*

	OLS analysis		LAD analysis	
Hours	R^2	Probability	$\Re$	Probability
4	0.2539	0.2681	0.1692	0.0378
40	0.5786	0.0153	0.2762	0.0034
$\lvert\Delta\rvert$	0.3247	0.2528	0.1070	0.0344

in Table 10.23. The value of 40 h of housework by the husband in family "L" is, by any definition, an extreme value. It is readily apparent that the extreme value of 40 h had a profound impact on the results of the OLS analysis. The OLS multiple correlation coefficient more than doubled from $R^2 = 0.2539$ to $R^2 = 0.5786$, yielding a difference between the two OLS multiple correlation coefficients of

$$\Delta_{R^2} = 0.5786 - 0.2539 = 0.3247 ,$$

and the corresponding exact probability value decreased from $P = 0.2681$ to $P = 0.0153$, yielding a difference between the two OLS probability values of

$$\Delta_P = 0.2681 - 0.0153 = 0.2528 .$$

The impact of 40 h of housework on the LAD analysis is more modest with the LAD chance-corrected measure of agreement increasing only slightly from $\Re = 0.1692$ to $\Re = 0.2762$, yielding a difference between the two LAD multiple correlation coefficients of

$$\Delta_{\Re} = 0.2762 - 0.1692 = 0.1070 ,$$

and the exact probability value decreasing from $P = 0.0378$ to $P = 0.0034$, yielding a difference between the two LAD probability values of only

$$\Delta_P = 0.0378 - 0.0034 = 0.0344 .$$

10.8 Summary

Under the Neyman–Pearson population model of statistical inference, this chapter examined product-moment linear correlation and regression, including both simple and multiple linear correlation and regression. The conventional measure of effect

size for simple OLS correlation and regression is Pearson's r_{xy}^2. Under the Fisher–Pitman permutation model of statistical inference, test statistics δ and associated measure of effect size $\Re$ were developed and illustrated for simple correlation and regression.

As in previous chapters, six examples illustrated statistics δ and $\Re$ for measures of linear correlation and regression. In the first example, a small sample of $N = 4$ bivariate observations was utilized to describe and simplify the calculation of statistics δ and $\Re$ for linear correlation and regression. The second example developed the permutation-based, chance-corrected measure of effect size, $\Re$, and related the permutation measure to Pearson's r_{xy}^2 measure of effect size. The third example with $N = 10$ bivariate observations illustrated the effects of extreme values on both ordinary least squares (OLS) regression based on squared Euclidean scaling with $v = 2$ and least absolute deviation (LAD) regression based on ordinary Euclidean scaling with $v = 1$. The fourth example with $N = 12$ bivariate observations compared exact and Monte Carlo probability procedures. A Monte Carlo permutation procedure was shown to be an accurate and efficient alternative to the calculation of an exact probability value, provided the probability value is not too small. The fifth example with $N = 11$ bivariate rank scores applied permutation statistical methods to rank-score correlation data, comparing permutation statistical methods to Spearman's rank-order correlation coefficient, Kendall's rank-order correlation coefficient, and Spearman's footrule correlation coefficient. The sixth example extended statistics δ and $\Re$ to multivariate correlation data. An example with $N = 12$ multivariate observations was analyzed with both OLS and LAD regression. A final example containing an extreme value provided a comparison of the two regression models when extreme values occur.

Chapter 11 concludes the presentation of permutation statistical methods with analyses of contingency tables. Six examples illustrate various permutation procedures applied to the analysis of contingency tables. The first example is devoted to goodness-of-fit tests. The second example considers contingency tables in which two nominal-level (categorical) variables have been cross-classified. The third example considers contingency tables in which two ordinal-level (ranked) variables have been cross-classified. The fourth example considers contingency tables in which one nominal-level variable and one ordinal-level variable have been cross-classified. The fifth example considers contingency tables in which one nominal-level variable and one interval-level variable have been cross-classified. The sixth example considers contingency tables in which one ordinal-level variable and one interval-level variable have been cross-classified.

References

1. Barrodale, I., Roberts, F.D.K.: A improved algorithm for discrete ℓ_1 linear approximation. J. Numer. Anal. **10**, 839–848 (1973)
2. Barrodale, I., Roberts, F.D.K.: Solution of an overdetermined system of equations in the ℓ_1 norm. Commun. ACM **17**, 319–320 (1974)

3. Berry, K.J., Mielke, P.W.: Least sum of absolute deviations regression: distance, leverage, and influence. Percept. Motor Skill. **86**, 1063–1070 (1998)
4. Berry, K.J., Mielke, P.W.: A Monte Carlo investigation of the Fisher *Z* transformation for normal and nonnormal distributions. Psychol. Rep. **87**, 1101–1114 (2000)
5. Hotelling, H., Pabst, M.R.: Rank correlation and tests of significance involving no assumption of normality. Ann. Math. Stat. **7**, 29–43 (1936)
6. Johnston, J.E., Berry, K.J., Mielke, P.W.: Permutation tests: precision in estimating probability values. Percept. Motor Skill. **105**, 915–920 (2007)