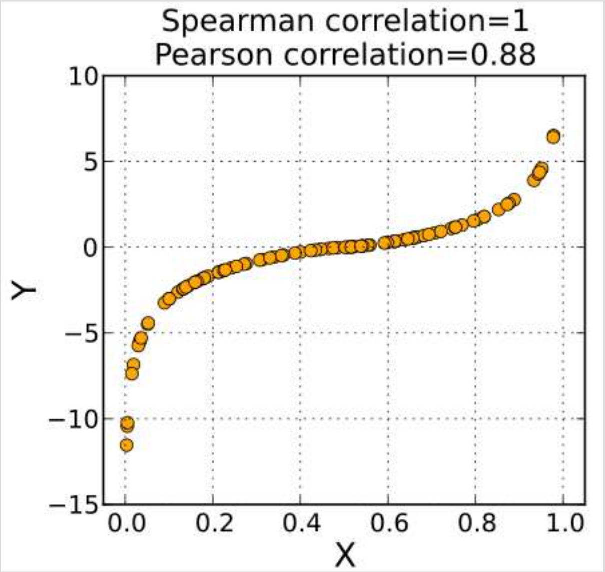
Wikipedia
The Free Encyclopedia

Spearman's rank correlation coefficient

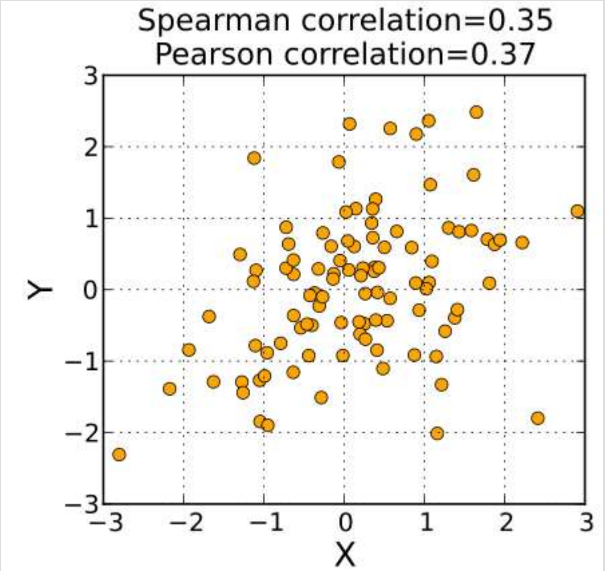
In statistics, **Spearman's rank correlation coefficient** or **Spearman's ρ**, named after Charles Spearman[1] and often denoted by the Greek letter ρ (rho) or as r_s, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.



A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater x values than that of a given data point will have greater y values as well. In contrast, this does not give a perfect Pearson correlation.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.



When the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.

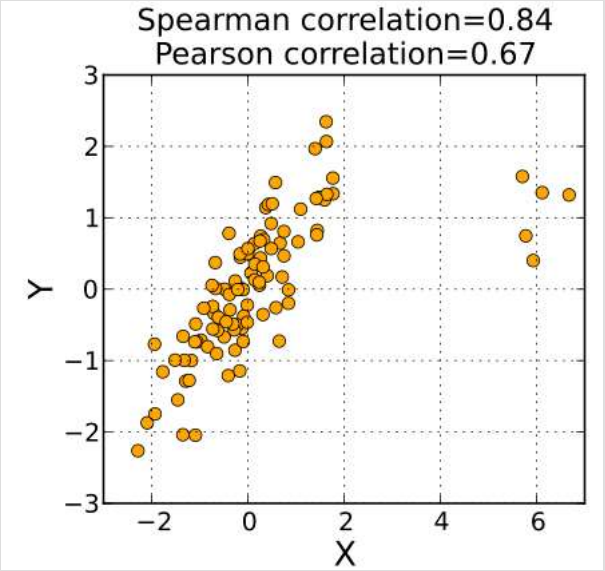
Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.[2][3] Both Spearman's ρ and Kendall's τ can be formulated as special cases of a more general correlation coefficient.

Applications

The coefficient can be used to determine how well data fits a model,[4] or to determine the similarity of text documents.[5]

Definition and calculation

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables.[6]



The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's ρ limits the outlier to the value of its rank.

For a sample of size n, the n raw scores X_i, Y_i are converted to ranks $\operatorname{R}(X_i), \operatorname{R}(Y_i)$, and r_s is computed as

$$r_s = \rho_{\operatorname{R}(X),\operatorname{R}(Y)} = \frac{\operatorname{cov}(\operatorname{R}(X), \operatorname{R}(Y))}{\sigma_{\operatorname{R}(X)} \sigma_{\operatorname{R}(Y)}},$$

where

ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,
$\operatorname{cov}(\operatorname{R}(X), \operatorname{R}(Y))$ is the covariance of the rank variables,
$\sigma_{\operatorname{R}(X)}$ and $\sigma_{\operatorname{R}(Y)}$ are the standard deviations of the rank variables.

Only if all n ranks are *distinct integers*, it can be computed using the popular formula

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where

$d_i = \operatorname{R}(X_i) - \operatorname{R}(Y_i)$ is the difference between the two ranks of each observation,
n is the number of observations.

[Proof] [show]

Consider a bivariate sample $(x_i, y_i),\ i = 1\ldots, n$ with corresponding ranks $(R(X_i), R(Y_i)) = (R_i, S_i)$. Then the Spearman correlation coefficient of x, y is

$$r_s = \frac{\sum_{i=1}^n R_i S_i - n\overline{R}\,\overline{S}}{\sigma_R \sigma_S},$$

where, as usual, $\overline{R} = \frac{1}{n}\sum_{i=1}^n R_i$, $\overline{S} = \frac{1}{n}\sum_{i=1}^n S_i$, $\sigma_R^2 = \frac{1}{n}\sum_{i=1}^n (R_i - \overline{R})^2$, and $\sigma_S^2 = \frac{1}{n}\sum_{i=1}^n (S_i - \overline{S})^2$,

We shall show that r_s can be expressed purely in terms of $d_i := R_i - S_i$, provided we assume that there be no ties within each sample.

Under this assumption, we have that R, S can be viewed as random variables distributed like a uniformly distributed random variable, U, on $\{1, 2, \ldots, n\}$. Hence $\overline{R} = \overline{S} = \mathbb{E}[U]$ and $\sigma_R^2 = \sigma_S^2 = \mathrm{Var}(U) = \mathbb{E}[U^2] - \mathbb{E}[U]^2$, where $\mathbb{E}[U] = \frac{1}{n}\sum_{i=1}^n i = \frac{(n+1)}{2}$, $\mathbb{E}[U^2] = \frac{1}{n}\sum_{i=1}^n i^2 = \frac{(n+1)(2n+1)}{6}$, and thus $\mathrm{Var}(U) = \frac{(n+1)(2n+1)}{6} - \left(\frac{(n+1)}{2}\right)^2 = \frac{n^2-1}{12}$. (These sums can be computed using the formulas for the triangular number and Square pyramidal number, or basic summation results from discrete mathematics.)

Observe now that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n R_i S_i - \overline{R}\overline{S} &= \frac{1}{n}\sum_{i=1}^n \frac{1}{2}(R_i^2 + S_i^2 - d_i^2) - \overline{R}^2 \\
&= \frac{1}{2}\frac{1}{n}\sum_{i=1}^n R_i^2 + \frac{1}{2}\frac{1}{n}\sum_{i=1}^n S_i^2 - \frac{1}{2n}\sum_{i=1}^n d_i^2 - \overline{R}^2 \\
&= (\frac{1}{n}\sum_{i=1}^n R_i^2 - \overline{R}^2) - \frac{1}{2n}\sum_{i=1}^n d_i^2 \\
&= \sigma_R^2 - \frac{1}{2n}\sum_{i=1}^n d_i^2 \\
&= \sigma_R\sigma_S - \frac{1}{2n}\sum_{i=1}^n d_i^2
\end{aligned}$$

Putting this all together thus yields

$$r_s = \frac{\sigma_R\sigma_S - \frac{1}{2n}\sum_{i=1}^n d_i^2}{\sigma_R\sigma_S} = 1 - \frac{\sum_{i=1}^n d_i^2}{2n \cdot \frac{n^2-1}{12}} = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)}.$$

Identical values are usually[7] each assigned fractional ranks equal to the average of their positions in the ascending order of the values, which is equivalent to averaging over all possible permutations.

If ties are present in the data set, the simplified formula above yields incorrect results: Only if in both variables all ranks are distinct, then $\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)} = \mathrm{Var}\left(\mathrm{R}(X)\right) = \mathrm{Var}\left(\mathrm{R}(Y)\right) = (n^2-1)/12$ (calculated according to biased variance). The first equation — normalizing by the standard deviation — may be used even when ranks are normalized to [0, 1] ("relative ranks") because it is insensitive both to translation and linear scaling.

The simplified method should also not be used in cases where the data set is truncated; that is, when the Spearman's correlation coefficient is desired for the top X records (whether by pre-change rank or post-change rank, or both), the user should use the Pearson correlation coefficient formula given above.[8]

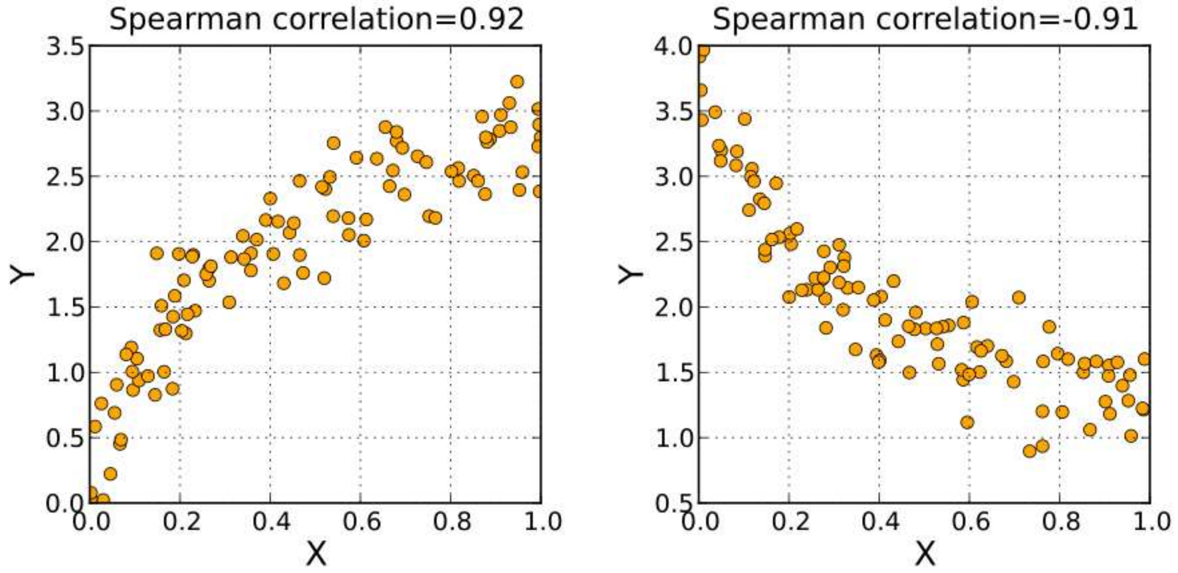
Related quantities

There are several other numerical measures that quantify the extent of statistical dependence between pairs of observations. The most common of these is the Pearson product-moment correlation coefficient, which is a similar correlation method to Spearman's rank, that measures the “linear” relationships between the raw numbers rather than between their ranks.

An alternative name for the Spearman rank correlation is the “grade correlation”;[9] in this, the “rank” of an observation is replaced by the “grade”. In continuous distributions, the grade of an observation is, by convention, always one half less than the rank, and hence the grade and rank correlations are the same in this case. More generally, the “grade” of an observation is proportional to an estimate of the fraction of a population less than a given value, with the half-observation adjustment at observed values. Thus this corresponds to one possible treatment of tied ranks. While unusual, the term “grade correlation” is still in use.[10]

Interpretation

Positive and negative Spearman rank correlations



A positive Spearman correlation coefficient corresponds to an increasing monotonic trend between *X* and *Y*.

A negative Spearman correlation coefficient corresponds to a decreasing monotonic trend between *X* and *Y*.

The sign of the Spearman correlation indicates the direction of association between X (the independent variable) and Y (the dependent variable). If Y tends to increase when X increases, the Spearman correlation coefficient is positive. If Y tends to decrease when X increases, the Spearman correlation coefficient is negative. A Spearman correlation of zero indicates that there is no tendency for Y to either increase or decrease when X increases. The Spearman correlation increases in magnitude as X and Y become closer to being perfectly monotone functions of each other. When X and Y are perfectly monotonically related, the Spearman correlation coefficient becomes 1. A perfectly monotone increasing relationship implies that for any two pairs of data values X_i, Y_i and X_j, Y_j, that $X_i - X_j$ and $Y_i - Y_j$ always have the same sign. A perfectly monotone decreasing relationship implies that these differences always have opposite signs.

The Spearman correlation coefficient is often described as being "nonparametric". This can have two meanings. First, a perfect Spearman correlation results when X and Y are related by any monotonic function. Contrast this with the Pearson correlation, which only gives a perfect value when X and Y are related by a *linear* function. The other sense in which the Spearman correlation is nonparametric is that its exact sampling distribution can be obtained without requiring knowledge (i.e., knowing the parameters) of the joint probability distribution of X and Y.

Example

In this example, the arbitrary raw data in the table below is used to calculate the correlation between the IQ of a person with the number of hours spent in front of TV per week [fictitious values used].

IQ, X_i	**Hours of TV per week**, Y_i
106	7
100	27
86	2
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Firstly, evaluate d_i^2. To do so use the following steps, reflected in the table below.

1. Sort the data by the first column (X_i). Create a new column x_i and assign it the ranked values 1, 2, 3, ..., *n*.
2. Next, sort the augmented (with x_i) data by the second column (Y_i). Create a fourth column y_i and similarly assign it the ranked values 1, 2, 3, ..., *n*.
3. Create a fifth column d_i to hold the differences between the two rank columns (x_i and y_i).
4. Create one final column d_i^2 to hold the value of column d_i squared.

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

With d_i^2 found, add them to find $\sum d_i^2 = 194$. The value of n is 10. These values can now be substituted back into the equation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

to give

$$\rho = 1 - \frac{6 \times 194}{10(10^2 - 1)},$$

which evaluates to $\rho = -29/165 = -0.175757575...$ with a p-value = 0.627188 (using the t-distribution).

That the value is close to zero shows that the correlation between IQ and hours spent watching TV is very low, although the negative value suggests that the longer the time spent watching television the lower the IQ. In the case of ties in the original values, this formula should not be used; instead, the Pearson correlation coefficient should be calculated on the ranks (where ties are given ranks, as described above).

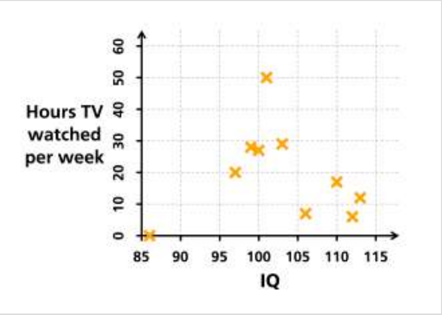


Chart of the data presented. It can be seen that there might be a negative correlation, but that the relationship does not appear definitive.

Confidence intervals

Confidence intervals for Spearman's ρ can be easily obtained using the Jackknife Euclidean likelihood approach in de Carvalho and Marques (2012).[11] The confidence interval with level α is based on a Wilks' theorem given in the latter paper, and is given by

$$\left\{ \theta : \frac{\{\sum_{i=1}^{n}(Z_i - \theta)\}^2}{\sum_{i=1}^{n}(Z_i - \theta)^2} \leq \chi_{1,\alpha}^2 \right\},$$

where $\chi^2_{1,\alpha}$ is the α quantile of a chi-square distribution with one degree of freedom, and the Z_i are jackknife pseudo-values. This approach is implemented in the R package spearmanCI (https://cran.r-project.org/web/packages/spearmanCI/index.html).

Determining significance

One approach to test whether an observed value of *ρ* is significantly different from zero (*r* will always maintain −1 ≤ *r* ≤ 1) is to calculate the probability that it would be greater than or equal to the observed *r*, given the null hypothesis, by using a permutation test. An advantage of this approach is that it automatically takes into account the number of tied data values in the sample and the way they are treated in computing the rank correlation.

Another approach parallels the use of the Fisher transformation in the case of the Pearson product-moment correlation coefficient. That is, confidence intervals and hypothesis tests relating to the population value *ρ* can be carried out using the Fisher transformation:

$$F(r) = \frac{1}{2} \ln \frac{1+r}{1-r} = \operatorname{artanh} r.$$

If *F*(*r*) is the Fisher transformation of *r*, the sample Spearman rank correlation coefficient, and *n* is the sample size, then

$$z = \sqrt{\frac{n-3}{1.06}} F(r)$$

is a *z*-score for *r*, which approximately follows a standard normal distribution under the null hypothesis of statistical independence ($\rho = 0$).[12][13]

One can also test for significance using

$$t = r \sqrt{\frac{n-2}{1-r^2}},$$

which is distributed approximately as Student's *t*-distribution with $n-2$ degrees of freedom under the null hypothesis.[14] A justification for this result relies on a permutation argument.[15]

A generalization of the Spearman coefficient is useful in the situation where there are three or more conditions, a number of subjects are all observed in each of them, and it is predicted that the observations will have a particular order. For example, a number of subjects might each be given three trials at the same task, and it is predicted that performance will improve from trial to trial. A test of the significance of the trend between conditions in this situation was developed by E. B. Page[16] and is usually referred to as Page's trend test for ordered alternatives.

Correspondence analysis based on Spearman's *ρ*

Classic correspondence analysis is a statistical method that gives a score to every value of two nominal variables. In this way the Pearson correlation coefficient between them is maximized.