Spearman's rank correlation coefficient

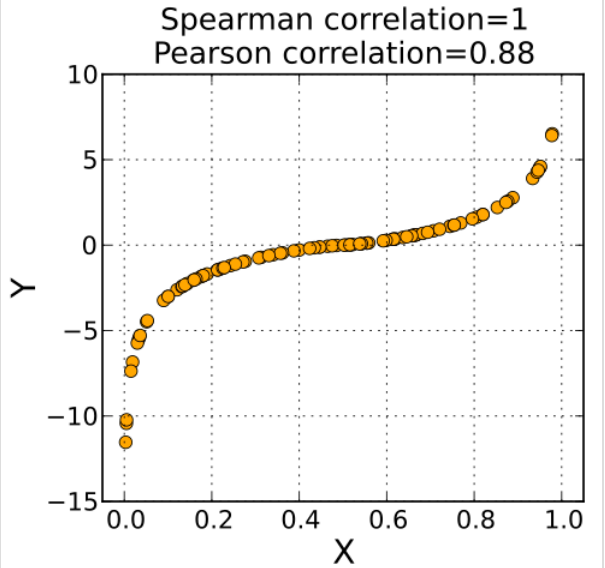
In statistics, **Spearman's rank correlation coefficient** or **Spearman's ρ** is a number ranging from -1 to 1 that indicates how strongly two sets of ranks are correlated. It could be used in a situation where one only has ranked data, such as a tally of gold, silver, and bronze medals. If a statistician wanted to know whether people who are high ranking in sprinting are also high ranking in long-distance running, they would use a Spearman rank correlation coefficient.

The coefficient is named after Charles Spearman[1] and often denoted by the Greek letter ρ (rho) or as r_s. It 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.

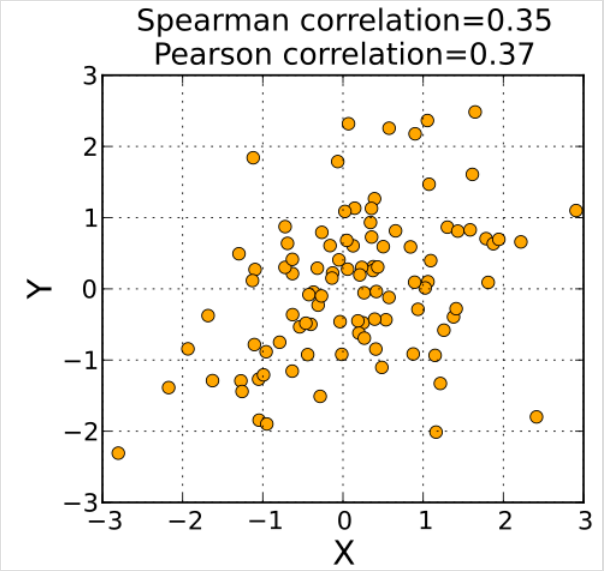
The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data 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.

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.



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.



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

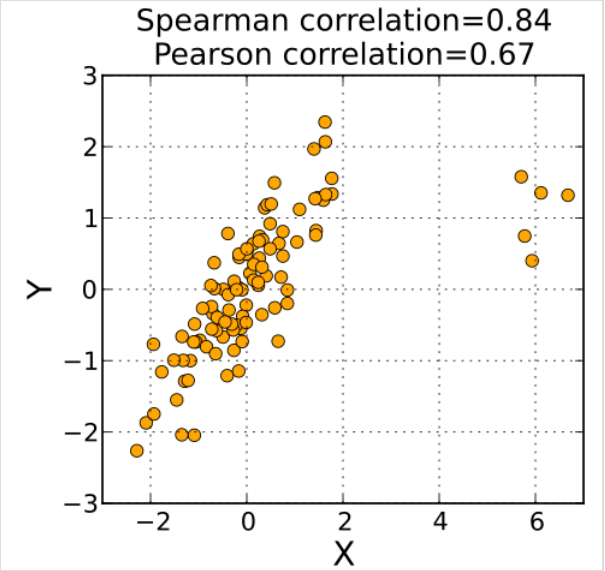
Applications

The coefficient can be used to determine how well data fits a model,[4] like when determining 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]

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



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.

$$r_s = \rho\left[\,\mathrm{R}[X], \mathrm{R}[Y]\,\right] = \frac{\operatorname{cov}\left[\,\mathrm{R}[X], \mathrm{R}[Y]\,\right]}{\sigma_{\mathrm{R}[X]}\ \sigma_{\mathrm{R}[Y]}},$$

where

ρ denotes the conventional Pearson correlation coefficient operator, but applied to the rank variables,
$\operatorname{cov}\left[\,\mathrm{R}[X], \mathrm{R}[Y]\,\right]$ is the covariance of the rank variables,
$\sigma_{\mathrm{R}[X]}$ and $\sigma_{\mathrm{R}[Y]}$ are the standard deviations of the rank variables.

Only when all n ranks are *distinct integers* (no ties), it can be computed using the popular formula

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

where

$d_i \equiv \mathrm{R}[X_i] - \mathrm{R}[Y_i]$ is the difference between the two ranks of each observation,
n is the number of observations.

[Proof]

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 $\sigma_{\mathrm{R}[X]}\ \sigma_{\mathrm{R}[Y]} = \operatorname{var}\left[\,\mathrm{R}[X]\,\right] = \operatorname{var}\left[\,\mathrm{R}[Y]\,\right] =$

$\frac{1}{12}\left(n^2-1\right)$ (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.

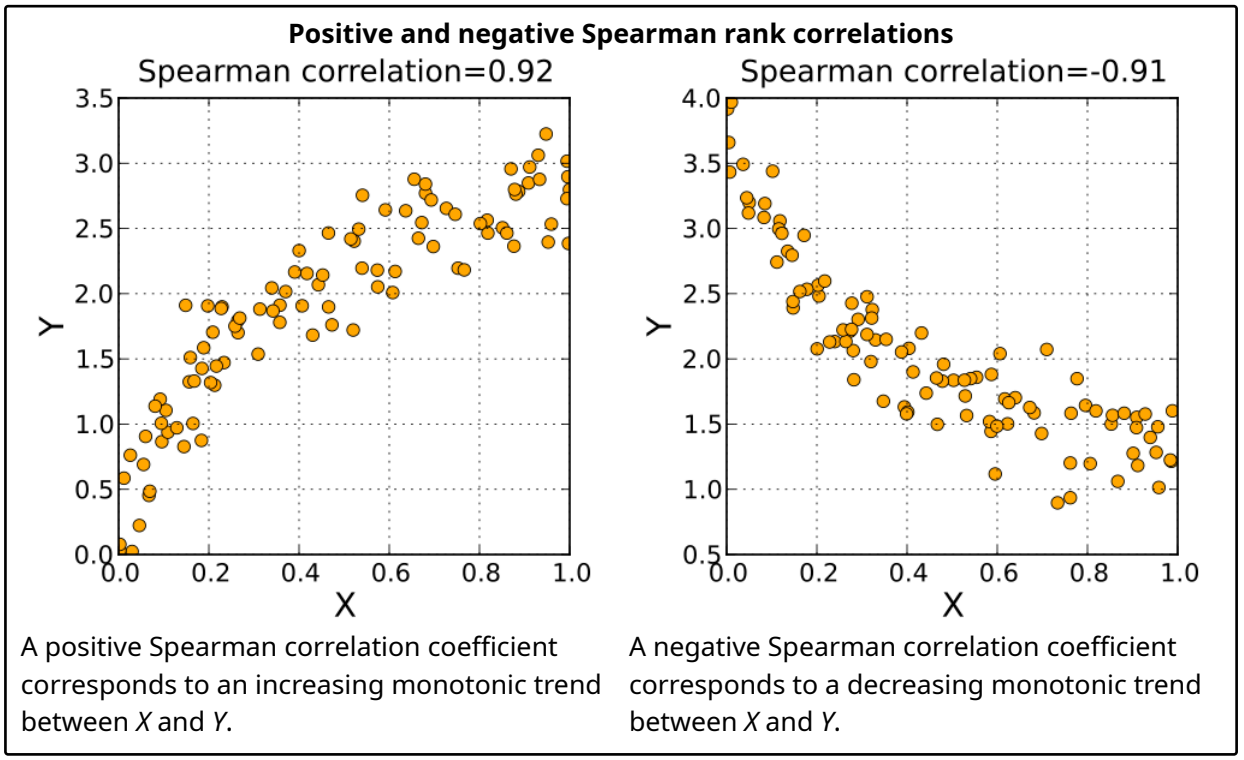
The simplified method should also not be used in cases where the data set is truncated; that is, when the Spearman's correlation coefficient is desired for the top *X* records (whether by pre-change rank or post-change rank, or both), the user should use the Pearson correlation coefficient formula given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



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 monotonic functions of each other. When X and Y are perfectly monotonically related, the Spearman correlation coefficient becomes 1. A perfectly monotonic 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 monotonic 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.

The strength of correlations also are an important consideration for interpretation of correlation analyses but descriptions of the strength of correlation are not universally accepted. A small correlation coefficient calculated for a very large sample may be statistically significant without providing a quantitative relation between variables. Similarly, a large correlation coefficient calculated with a small sample size may indicate that a quantitative equation between variables may be possible even though the correlation coefficient is not statistically significant. Akoglu (2018)[11] noted the need for consistent descriptions of strength and provides different correlation-strength descriptors for psychology, politics, and medicine that have different descriptors and thresholds. Because there is a general lack of consensus on correlation strength, Granato (2014) [12] defined operational definitions for the absolute values of correlation coefficients as weak (less than 0.5), moderate (greater than or equal to 0.5 and less than 0.75), semi-strong (greater than or equal to 0.75 and less than 0.85), and strong (greater than or equal to 0.85) for use in hydrologic analyses of stormwater treatment statistics. Similarly, Schober and others (2018)[13] note that "...cutoff points are arbitrary and inconsistent and should be used judiciously..." Despite their warning Schober and others (2018) provide a table indicating that correlations less than or equal to 0.1 are negligible; correlations greater than 0.1 and less than or equal to 0.39 are weak; correlations greater than 0.39 and less than or equal to 0.69 are moderate; correlations greater than 0.69 and less than or equal to 0.89 are strong; and correlations greater than 0.89 are very strong. Descriptions of strength indicate the potential to develop quantitative relations among variables but these descriptors, as with the correlation coefficients themselves, do not indicate causation.

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).

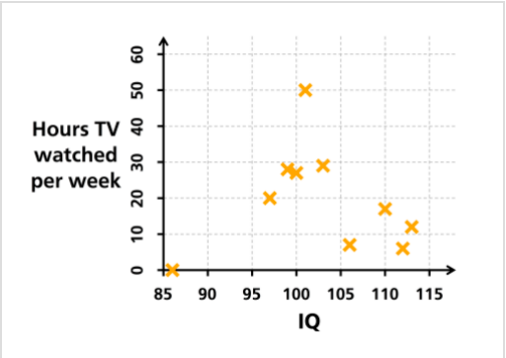


Chart of the data presented. It can be seen that there might be a negative correlation, but that the relationship does not appear definitive.

Confidence intervals

Confidence intervals for Spearman's ρ can be easily obtained using the Jackknife Euclidean likelihood approach in de Carvalho and Marques (2012).[14] The confidence interval with level α is based on a Wilks' theorem given in the latter paper, and is given by

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

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 $-1 \leq r \leq 1$) 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:

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

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 ($\rho = 0$).[15][16]

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.[17] A justification for this result relies on a permutation argument.[18]

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[19] 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.

There exists an equivalent of this method, called grade correspondence analysis, which maximizes Spearman's ρ or Kendall's τ.[20]

Approximating Spearman's ρ from a stream

There are two existing approaches to approximating the Spearman's rank correlation coefficient from streaming data.[21][22] The first approach[21] involves coarsening the joint distribution of (X, Y). For continuous X, Y values: m_1, m_2 cutpoints are selected for X and Y respectively, discretizing these random variables. Default cutpoints are added at $-\infty$ and ∞. A count matrix of size $(m_1 + 1) \times (m_2 + 1)$, denoted M, is then constructed where $M[i, j]$ stores the number of observations that fall into the two-dimensional cell indexed by (i, j). For streaming data, when a new observation arrives, the appropriate $M[i, j]$ element is incremented. The Spearman's rank correlation can then be computed, based on the count matrix M, using linear algebra operations (Algorithm 2[21]). Note that for discrete random variables, no discretization procedure is necessary. This method is applicable to

stationary streaming data as well as large data sets. For non-stationary streaming data, where the Spearman's rank correlation coefficient may change over time, the same procedure can be applied, but to a moving window of observations. When using a moving window, memory requirements grow linearly with chosen window size.

The second approach to approximating the Spearman's rank correlation coefficient from streaming data involves the use of Hermite series based estimators.[22] These estimators, based on Hermite polynomials, allow sequential estimation of the probability density function and cumulative distribution function in univariate and bivariate cases. Bivariate Hermite series density estimators and univariate Hermite series based cumulative distribution function estimators are plugged into a large sample version of the Spearman's rank correlation coefficient estimator, to give a sequential Spearman's correlation estimator. This estimator is phrased in terms of linear algebra operations for computational efficiency (equation (8) and algorithm 1 and 2[22]). These algorithms are only applicable to continuous random variable data, but have certain advantages over the count matrix approach in this setting. The first advantage is improved accuracy when applied to large numbers of observations. The second advantage is that the Spearman's rank correlation coefficient can be computed on non-stationary streams without relying on a moving window. Instead, the Hermite series based estimator uses an exponential weighting scheme to track time-varying Spearman's rank correlation from streaming data, which has constant memory requirements with respect to "effective" moving window size. A software implementation of these Hermite series based algorithms exists [23] and is discussed in Software implementations.

Software implementations

- R's statistics base-package implements the test `cor.test(x, y, method = "spearman")` (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/cor.test.html) in its "stats" package (also `cor(x, y, method = "spearman")` will work). The package spearmanCI (https://cran.r-project.org/web/packages/spearmanCI/index.html) computes confidence intervals. The package hermiter (https://cran.r-project.org/package=hermiter)[23] computes fast batch estimates of the Spearman correlation along with sequential estimates (i.e. estimates that are updated in an online/incremental manner as new observations are incorporated).
- Stata implementation: `spearman` *`varlist`* calculates all pairwise correlation coefficients for all variables in *varlist*.
- MATLAB implementation: `[r,p] = corr(x,y,'Type','Spearman')` where `r` is the Spearman's rank correlation coefficient, `p` is the p-value, and `x` and `y` are vectors.[24]
- Python has many different implementations of the spearman correlation statistic: it can be computed with the spearmanr (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html) function of the `scipy.stats` module, as well as with the `DataFrame.corr(method='spearman')` method from the pandas (https://pandas.pydata.org/docs/index.html) library, and the `corr(x, y, method='spearman')` function from the statistical package pingouin (https://pingouin-stats.org/index.html).

See also