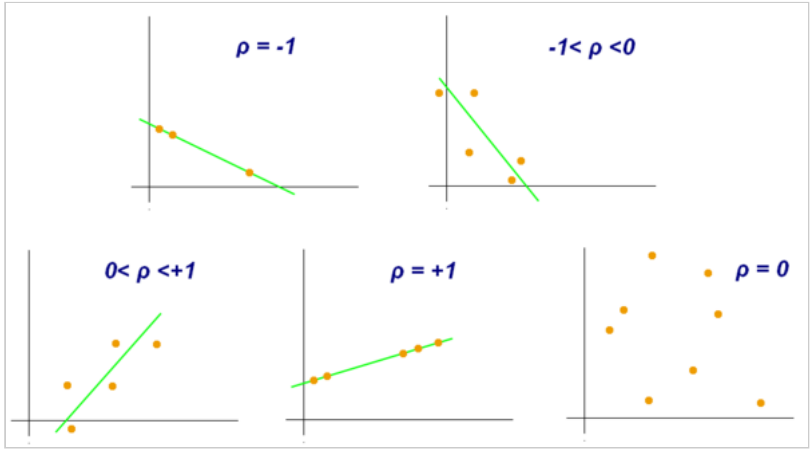
Pearson correlation coefficient

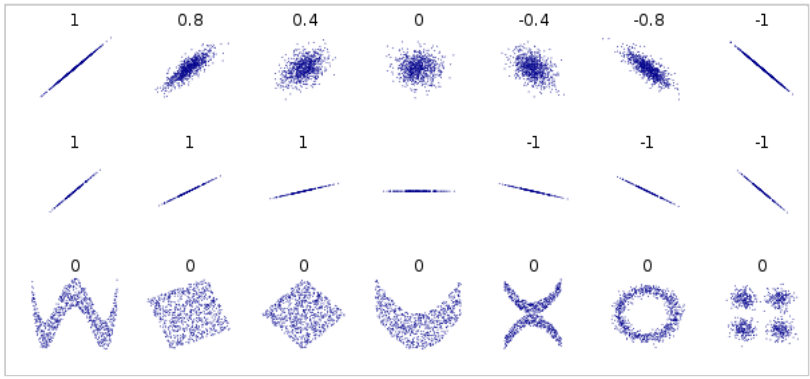
From Wikipedia, the free encyclopedia

In statistics, the **Pearson correlation coefficient** (**PCC**, pronounced /ˈpɪərsən/), also referred to as the **Pearson's *r***, **Pearson product-moment correlation coefficient** (**PPMCC**) or **bivariate correlation**,[1] is a measure of the linear correlation between two variables *X* and *Y*. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is widely used in the sciences. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.[2][3][4]

Contents



Examples of scatter diagrams with different values of correlation coefficient (ρ)



Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the non-linearity and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

Definition

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

For a population

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. The formula for ρ[5] is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

The formula for ρ can be expressed in terms of mean and expectation. Since

$$\operatorname{cov}(X,Y) = \operatorname{E}[(X-\mu_X)(Y-\mu_Y)]$$[5]

Then the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\operatorname{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where:

- cov and σ_X are defined as above
- μ_X is the mean of X
- E is the expectation.

The formula for ρ can be expressed in terms of uncentered moments. Since

- $\mu_X = \operatorname{E}[X]$
- $\mu_Y = \operatorname{E}[Y]$
- $\sigma_X^2 = \operatorname{E}[(X - \operatorname{E}[X])^2] = \operatorname{E}[X^2] - [\operatorname{E}[X]]^2$
- $\sigma_Y^2 = \operatorname{E}[(Y - \operatorname{E}[Y])^2] = \operatorname{E}[Y^2] - [\operatorname{E}[Y]]^2$
- $\operatorname{E}[(X-\mu_X)(Y-\mu_Y)] = \operatorname{E}[(X-\operatorname{E}[X])(Y-\operatorname{E}[Y])] = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y],$

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y]}{\sqrt{\operatorname{E}[X^2] - [\operatorname{E}[X]]^2}\ \sqrt{\operatorname{E}[Y^2] - [\operatorname{E}[Y]]^2}}.$$

For a sample

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter *r* and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for *r* by substituting estimates of the covariances and variances based on a sample into the formula above. So if we have one dataset $\{x_1,...,x_n\}$ containing *n* values and another dataset $\{y_1,...,y_n\}$ containing *n* values then that formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n, x_i, y_i are defined as above
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Rearranging gives us this formula for *r*:

$$r = r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

where:

- n, x_i, y_i are defined as above
- This formula suggests a convenient single-pass algorithm for calculating sample correlations, but, depending on the numbers involved, it can sometimes be numerically unstable.

Rearranging again gives us this[5] formula for *r*:

$$r = r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\ \sqrt{(\sum y_i^2 - n\bar{y}^2)}}.$$

where:

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above

An equivalent expression gives the formula for *r* as the mean of the products of the standard scores as follows:

$$r = r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above, and s_x, s_y are defined below
- $\left(\frac{x_i - \bar{x}}{s_x}\right)$ is the standard score (and analogously for the standard score of *y*)

Alternative formulae for *r* are also available. One can use the following formula for *r*:

$$r = r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

where:

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above and:
- $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (the sample standard deviation); and analogously for s_y

Practical issues

Under heavy noise conditions, extracting the correlation coefficient between two sets of stochastic variables is nontrivial, in particular where Canonical Correlation Analysis reports on degraded correlation values due to the heavy noise contributions. A generalization of the approach is given elsewhere.[6]

In case of missing data, Garren derived the maximum likelihood estimator.[7]

Mathematical properties

The absolute values of both the sample and population Pearson correlation coefficients are less than or equal to 1. Correlations equal to 1 or −1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation). The Pearson correlation coefficient is symmetric: corr(*X*,*Y*) = corr(*Y*,*X*).

A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may transform *X* to *a* + *bX* and transform *Y* to *c* + *dY*, where *a*, *b*, *c*, and *d* are constants with *b*, *d* > 0, without changing the correlation coefficient. (This holds for both the population and sample Pearson correlation coefficients.) Note that more general linear transformations do change the correlation: see § Decorrelation for an application of this.

Interpretation

The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line for which *Y* increases as *X* increases. A value of −1 implies that all data points lie on a line for which *Y* decreases as *X* increases. A value of 0 implies that there is no linear correlation between the variables.

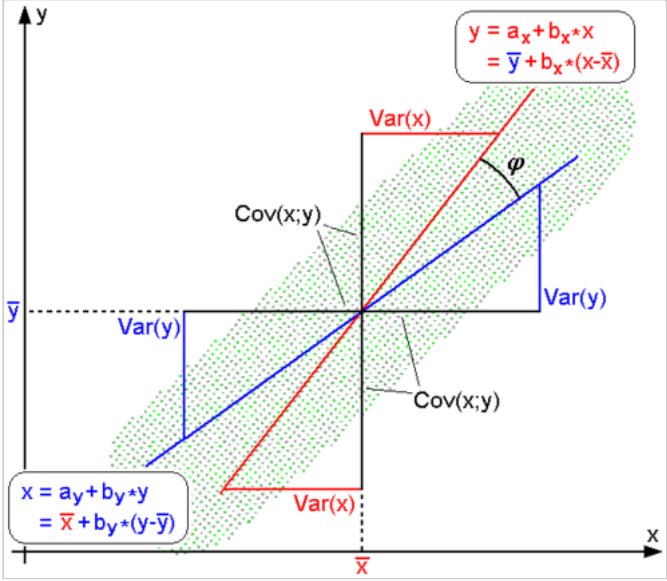
More generally, note that $(X_i - \overline{X})(Y_i - \overline{Y})$ is positive if and only if X_i and Y_i lie on the same side of their respective means. Thus the correlation coefficient is positive if X_i and Y_i tend to be simultaneously greater than, or simultaneously less than, their respective means. The correlation coefficient is negative (anti-correlation) if X_i and Y_i tend to lie on opposite sides of their respective means. Moreover, the stronger is either tendency, the larger is the absolute value of the correlation coefficient.

Geometric interpretation

For uncentered data, there is a relation between the correlation coefficient and the angle φ between the two regression lines, $y = g_x(x)$ and $x = g_y(y)$, obtained by regressing *y* on *x* and *x* on *y* respectively. (Here φ is measured within the first quadrant formed around the lines' intersection point if $r > 0$, or counterclockwise from the fourth to

the second quadrant if $r < 0$.) One can show[8] that if the standard deviations are equal, then $r = \sec\varphi - \tan\varphi$, where sec and tan are trigonometric functions.

For centered data (i.e., data which have been shifted by the sample means of their respective variables so as to have an average of zero for each variable), the correlation coefficient can also be viewed as the cosine of the angle θ between the two vectors of samples in N-dimensional space (for N samples of each variable)[9]:ch. 5 (as illustrated for a special case in the next paragraph).



Regression lines for $y = g_x(x)$ [red] and $x = g_y(y)$ [blue]

Both the uncentered (non-Pearson-compliant) and centered correlation coefficients can be determined for a dataset. As an example, suppose five countries are found to have gross national products of 1, 2, 3, 5, and 8 billion dollars, respectively. Suppose these same five countries (in the same order) are found to have 11%, 12%, 13%, 15%, and 18% poverty. Then let **x** and **y** be ordered 5-element vectors containing the above data: **x** = (1, 2, 3, 5, 8) and **y** = (0.11, 0.12, 0.13, 0.15, 0.18).

By the usual procedure for finding the angle θ between two vectors (see dot product), the *uncentered* correlation coefficient is:

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \frac{2.93}{\sqrt{103}\sqrt{0.0983}} = 0.920814711.$$

This uncentred correlation coefficient is identical with the cosine similarity. Note that the above data were deliberately chosen to be perfectly correlated: $y = 0.10 + 0.01\,x$. The Pearson correlation coefficient must therefore be exactly one. Centering the data (shifting **x** by E(**x**) = 3.8 and **y** by E(**y**) = 0.138) yields **x** = (−2.8, −1.8, −0.8, 1.2, 4.2) and **y** = (−0.028, −0.018, −0.008, 0.012, 0.042), from which

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \frac{0.308}{\sqrt{30.8}\sqrt{0.00308}} = 1 = \rho_{xy},$$

as expected.

Interpretation of the size of a correlation

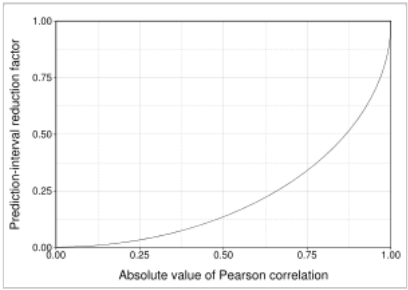
Several authors have offered guidelines for the interpretation of a correlation coefficient.[10][11] However, all such criteria are in some ways arbitrary.[11] The interpretation of a correlation coefficient depends on the context and purposes. A correlation of 0.8 may be very low if one is verifying a physical law using high-quality instruments, but may be regarded as very high in the social sciences where there may be a greater contribution from complicating factors.

Inference

Statistical inference based on Pearson's correlation coefficient often focuses on one of the following two aims:

- One aim is to test the null hypothesis that the true correlation coefficient ρ is equal to 0, based on the value of the sample correlation coefficient r.
- The other aim is to derive a confidence interval that, on repeated sampling, has a given probability of containing ρ.

We discuss methods of achieving one or both of these aims below.



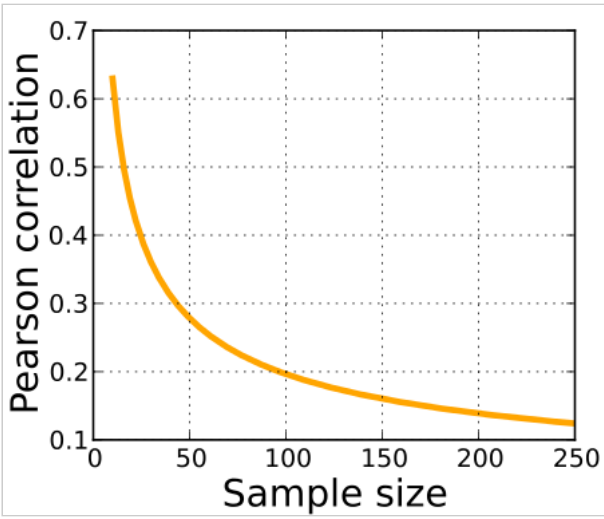
This figure gives a sense of how the usefulness of a Pearson correlation for predicting values varies with its magnitude. Given jointly normal X, Y with correlation ρ, $1-\sqrt{1-\rho^2}$ (plotted here as a function of ρ) is the factor by which a given prediction interval for Y may be reduced given the corresponding value of X. For example, if $\rho = .5$, then the 95% prediction interval of $Y|X$ will be about 13% smaller than the 95% prediction interval of Y.

Using a permutation test

Permutation tests provide a direct approach to performing hypothesis tests and constructing confidence intervals. A permutation test for Pearson's correlation coefficient involves the following two steps:

1. Using the original paired data (x_i, y_i), randomly redefine the pairs to create a new data set $(x_i, y_{i'})$, where the i' are a permutation of the set $\{1,...,n\}$. The permutation i' is selected randomly, with equal probabilities placed on all $n!$ possible permutations. This is equivalent to drawing the i' randomly "**without replacement**" from the set $\{1, ..., n\}$. A closely related and equally justified (bootstrapping) approach is to separately draw the i and the i' "**with replacement**" from $\{1, ..., n\}$;
2. Construct a correlation coefficient r from the randomized data.

To perform the permutation test, repeat steps (1) and (2) a large number of times. The p-value for the permutation test is the proportion of the r values generated in step (2) that are larger than the Pearson correlation coefficient that was calculated from the original data. Here "larger" can mean either that the value is larger in magnitude, or larger in signed value, depending on whether a two-sided or one-sided test is desired.



A graph showing the minimum absolute value of Pearson's correlation coefficient that is significantly different from zero at the 0.05 level, for a given sample size.

Using a bootstrap

The bootstrap can be used to construct confidence intervals for Pearson's correlation coefficient. In the "non-parametric" bootstrap, n pairs (x_i, y_i) are resampled "with replacement" from the observed set of n pairs, and the correlation coefficient r is calculated based on the resampled data. This process is repeated a large number of times, and the empirical distribution of the resampled r values are used to approximate the sampling distribution of the statistic. A 95% confidence interval for ρ can be defined as the interval spanning from the 2.5th to the 97.5th percentile of the resampled r values.

Testing using Student's *t*-distribution

For pairs from an uncorrelated bivariate normal distribution, the sampling distribution of a certain function of Pearson's correlation coefficient follows Student's t-distribution with degrees of freedom $n - 2$. Specifically, if the underlying variables have a bivariate normal distribution, the variable

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

has a Student's *t*-distribution in the null case (zero correlation).[12] This also holds approximately even if the observed values are non-normal, provided sample sizes are not very small.[13] For determining the critical values for *r* the inverse of this transformation is also needed:

$$r = \frac{t}{\sqrt{n-2+t^2}}.$$

Alternatively, large sample approaches can be used.

Another early paper[14] provides graphs and tables for general values of ρ, for small sample sizes, and discusses computational approaches.

Using the exact distribution

For data that follows a bivariate normal distribution, the exact density function *f*(*r*) for the sample correlation coefficient *r* of a normal bivariate is[15][16]

$$f(r) = \frac{(n-2)\,\Gamma(n-1)(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}}{\sqrt{2\pi}\,\Gamma\left(n-\frac{1}{2}\right)(1-\rho r)^{n-\frac{3}{2}}}\,{}_2\mathbf{F}_1\left(\frac{1}{2},\frac{1}{2};\frac{2n-1}{2};\frac{\rho r+1}{2}\right)$$

where:

- Γ is the gamma function,
- ${}_2\mathbf{F}_1(a,b;c;z)$ is the Gaussian hypergeometric function.

In the special case when $\rho = 0$, the exact density function *f(r)* can be written as:

$$f(r) = \frac{(1-r^2)^{\frac{n-4}{2}}}{\mathbf{B}\left(\frac{1}{2},\frac{n-2}{2}\right)},$$

where:

- $\mathbf{B}$ is the beta function, which is one way of writing the density of a Student's t-distribution, as above.

Using the Fisher transformation

In practice, confidence intervals and hypothesis tests relating to ρ are usually 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*, and *n* is the sample size, then *F*(*r*) approximately follows a normal distribution with

$$\text{mean} = F(\rho) = \operatorname{artanh}(\rho) \quad \text{and standard error} \quad \mathrm{SE} = \frac{1}{\sqrt{n-3}}.$$

Thus, a z-score is

$$z = \frac{x - \text{mean}}{\text{SE}} = [F(r) - F(\rho_0)]\sqrt{n-3}$$

under the null hypothesis of that $\rho = \rho_0$, given the assumption that the sample pairs are independent and identically distributed and follow a bivariate normal distribution. Thus an approximate p-value can be obtained from a normal probability table. For example, if $z = 2.2$ is observed and a two-sided p-value is desired to test the null hypothesis that $\rho = 0$, the p-value is $2 \cdot \Phi(-2.2) = 0.028$, where Φ is the standard normal cumulative distribution function.

To obtain a confidence interval for ρ, we first compute a confidence interval for $F(\rho)$:

$$100(1-\alpha)\%\text{CI} : \operatorname{artanh}(\rho) \in [\operatorname{artanh}(r) \pm z_{\alpha/2}SE]$$

The inverse Fisher transformation bring the interval back to the correlation scale.

$$100(1-\alpha)\%\text{CI} : \rho \in [\tanh(\operatorname{artanh}(r) - z_{\alpha/2}SE), \tanh(\operatorname{artanh}(r) + z_{\alpha/2}SE)]$$

For example, suppose we observe $r = 0.3$ with a sample size of n=50, and we wish to obtain a 95% confidence interval for ρ. The transformed value is $\operatorname{artanh}(r) = 0.30952$, so the confidence interval on the transformed scale is $0.30952 \pm 1.96/\sqrt{47}$, or (0.023624, 0.595415). Converting back to the correlation scale yields (0.024, 0.534).

In least squares regression analysis

The square of the sample correlation coefficient is typically denoted r^2 and is a special case of the coefficient of determination. In this case, it estimates the fraction of the variance in Y that is explained by X in a simple linear regression. So if we have the observed dataset $\{y_1,...,y_n\}$ and the fitted dataset $\{f_1,...,f_n\}$, and we denote the fitted dataset $\{f_1,...,f_n\}$ with $\{\hat{y}_1,...,\hat{y}_n\}$, then as a starting point the total variation in the Y_i around their average value can be decomposed as follows

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2,$$

where the $\hat{Y}_i$ are the fitted values from the regression analysis. This can be rearranged to give

$$1 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2} + \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}.$$

The two summands above are the fraction of variance in Y that is explained by X (right) and that is unexplained by X (left).

Next, we apply a property of least square regression models, that the sample covariance between $\hat{Y}_i$ and $Y_i - \hat{Y}_i$ is zero. Thus, the sample correlation coefficient between the observed and fitted response values in the regression can be written (calculation is under expectation, assumes Gaussian statistics)

$$
\begin{aligned}
r(Y, \hat{Y}) &= \frac{\sum_i (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\
&= \frac{\sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\
&= \frac{\sum_i [(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2]}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\
&= \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\
&= \sqrt{\frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}}.
\end{aligned}
$$

Thus

$$
r(Y, \hat{Y})^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}
$$

where

- $r(Y, \hat{Y})^2$ is the proportion of variance in Y explained by a linear function of X.

That equation can be written as:

$$
r(Y, \hat{Y})^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}}
$$

where

- SS_{reg} is the regression sum of squares, also called the explained sum of squares
- SS_{tot} is the total sum of squares (proportional to the variance of the data)
- $SS_{\text{reg}} = \sum_i (\hat{Y}_i - \bar{Y})^2$
- $SS_{\text{tot}} = \sum_i (Y_i - \bar{Y})^2$

Sensitivity to the data distribution

Existence

The population Pearson correlation coefficient is defined in terms of moments, and therefore exists for any bivariate probability distribution for which the population covariance is defined and the marginal population variances are defined and are non-zero. Some probability distributions such as the Cauchy distribution have undefined variance and hence ρ is not defined if X or Y follows such a distribution. In some practical applications,

such as those involving data suspected to follow a heavy-tailed distribution, this is an important consideration. However, the existence of the correlation coefficient is usually not a concern; for instance, if the range of the distribution is bounded, ρ is always defined.

Sample size

- If the sample size is moderate or large and the population is normal, then, in the case of the bivariate normal distribution, the sample correlation coefficient is the maximum likelihood estimate of the population correlation coefficient, and is asymptotically unbiased and efficient, which roughly means that it is impossible to construct a more accurate estimate than the sample correlation coefficient.
- If the sample size is large and the population is not normal, then the sample correlation coefficient remains approximately unbiased, but may not be efficient.
- If the sample size is large then sample correlation coefficient is a consistent estimator of the population correlation coefficient as long as the sample means, variances, and covariance are consistent (which is guaranteed when the law of large numbers can be applied).
- If the sample size is small then the sample correlation coefficient r is not an unbiased estimate of ρ.[5] The adjusted correlation coefficient must be used instead: see elsewhere in this article for the definition.

Robustness

Like many commonly used statistics, the sample statistic r is not robust,[17] so its value can be misleading if outliers are present.[18][19] Specifically, the PMCC is neither distributionally robust, nor outlier resistant[17] (see Robust statistics#Definition). Inspection of the scatterplot between X and Y will typically reveal a situation where lack of robustness might be an issue, and in such cases it may be advisable to use a robust measure of association. Note however that while most robust estimators of association measure statistical dependence in some way, they are generally not interpretable on the same scale as the Pearson correlation coefficient.

Statistical inference for Pearson's correlation coefficient is sensitive to the data distribution. Exact tests, and asymptotic tests based on the Fisher transformation can be applied if the data are approximately normally distributed, but may be misleading otherwise. In some situations, the bootstrap can be applied to construct confidence intervals, and permutation tests can be applied to carry out hypothesis tests. These non-parametric approaches may give more meaningful results in some situations where bivariate normality does not hold. However the standard versions of these approaches rely on exchangeability of the data, meaning that there is no ordering or grouping of the data pairs being analyzed that might affect the behavior of the correlation estimate.

A stratified analysis is one way to either accommodate a lack of bivariate normality, or to isolate the correlation resulting from one factor while controlling for another. If W represents cluster membership or another factor that it is desirable to control, we can stratify the data based on the value of W, then calculate a correlation coefficient within each stratum. The stratum-level estimates can then be combined to estimate the overall correlation while controlling for W.[20]

Variants

Variations of the correlation coefficient can be calculated for different purposes. Here are some examples.

Adjusted correlation coefficient

The sample correlation coefficient r is not an unbiased estimate of ρ. For data that follows a bivariate normal distribution, the expectation $E(r)$ for the sample correlation coefficient r of a normal bivariate is[21]

$$\mathbf{E}[r] = \rho - \frac{\rho\left(1-\rho^2\right)}{2n} + \cdots, \quad \text{therefore } r \text{ is a biased estimator of } \boldsymbol{\rho}.$$

The unique minimum variance unbiased estimator r_{adj} is given by[22]

$$(1) \qquad r_{\mathrm{adj}} = r\,{}_2\mathbf{F}_1\left(\frac{1}{2}, \frac{1}{2}; \frac{n-1}{2}; 1-r^2\right),$$

where:

- r, n are defined as above,
- ${}_2\mathbf{F}_1(a, b; c; z)$ is the Gaussian hypergeometric function.

An approximately unbiased estimator r_{adj} can be obtained by truncating $\mathrm{E}[r]$ and solving this truncated equation:

$$(2) \qquad r = \mathbf{E}[r] = r_{\mathrm{adj}} - \frac{r_{\mathrm{adj}}\left(1 - r_{\mathrm{adj}}^2\right)}{2n}.$$

The solution to equation (2) is:

$$(3) \qquad r_{adj} = r\left[1 + \frac{1-r^2}{2n}\right],$$

where in (3):

- r, n are defined as above,
- r_{adj} is a suboptimal estimator,
- r_{adj} can also be obtained by maximizing $\log(f(r))$,
- r_{adj} has minimum variance for large values of n,
- r_{adj} has a bias of order $1/(n-1)$.

Another proposed[5] adjusted correlation coefficient is:

$$r_{\mathrm{adj}} = \sqrt{1 - \frac{(1-r^2)(n-1)}{(n-2)}}.$$

Note that $r_{\mathrm{adj}} \approx r$ for large values of n.

Weighted correlation coefficient

Suppose observations to be correlated have differing degrees of importance that can be expressed with a weight vector w. To calculate the correlation between vectors x and y with the weight vector w (all of length n),[23][24]

- Weighted mean:

$$\mathrm{m}(x; w) = \frac{\sum_i w_i x_i}{\sum_i w_i}.$$

- Weighted covariance

$$\operatorname{cov}(x, y; w) = \frac{\sum_i w_i (x_i - \operatorname{m}(x; w))(y_i - \operatorname{m}(y; w))}{\sum_i w_i}.$$

- Weighted correlation

$$\operatorname{corr}(x, y; w) = \frac{\operatorname{cov}(x, y; w)}{\sqrt{\operatorname{cov}(x, x; w) \operatorname{cov}(y, y; w)}}.$$

Reflective correlation coefficient

The reflective correlation is a variant of Pearson's correlation in which the data are not centered around their mean values. The population reflective correlation is

$$\operatorname{Corr}_r(X, Y) = \frac{E[XY]}{\sqrt{EX^2 \cdot EY^2}}.$$

The reflective correlation is symmetric, but it is not invariant under translation:

$$\operatorname{Corr}_r(X, Y) = \operatorname{Corr}_r(Y, X) = \operatorname{Corr}_r(X, bY) \neq \operatorname{Corr}_r(X, a + bY), \quad a \neq 0, b > 0.$$

The sample reflective correlation is

$$rr_{xy} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}.$$

The weighted version of the sample reflective correlation is

$$rr_{xy,w} = \frac{\sum w_i x_i y_i}{\sqrt{(\sum w_i x_i^2)(\sum w_i y_i^2)}}.$$

Scaled correlation coefficient

Scaled correlation is a variant of Pearson's correlation in which the range of the data is restricted intentionally and in a controlled manner to reveal correlations between fast components in time series.[25] Scaled correlation is defined as average correlation across short segments of data.

Let K be the number of segments that can fit into the total length of the signal T for a given scale s:

$$K = \operatorname{round}\left(\frac{T}{s}\right).$$

The scaled correlation across the entire signals $\bar{r}_s$ is then computed as

$$\bar{r}_s = \frac{1}{K} \sum_{k=1}^{K} r_k,$$

where r_k is Pearson's coefficient of correlation for segment k.

By choosing the parameter s, the range of values is reduced and the correlations on long time scale are filtered out, only the correlations on short time scales being revealed. Thus, the contributions of slow components are removed and those of fast components are retained.

Pearson's distance

A distance metric for two variables X and Y known as *Pearson's distance* can be defined from their correlation coefficient as[26]

$$d_{X,Y} = 1 - \rho_{X,Y}.$$

Considering that the Pearson correlation coefficient falls between [−1, 1], the Pearson distance lies in [0, 2].

Circular correlation coefficient

For variables X = {x_1,...,x_n} and Y = {y_1,...,y_n} that are defined on the unit circle [0, 2π), it is possible to define a circular analog of Pearson's coefficient.[27] This is done by transforming data points in X and Y with a sine function such that the correlation coefficient is given as:

$$r_{\text{circular}} = \frac{\sum_{i=1}^{n} \sin(x_i - \bar{x}) \sin(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} \sin(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} \sin(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the circular means of *X* and *Y*. This measure can be useful in fields like meteorology where the angular direction of data is important.

Partial correlation

If a population or data-set is characterized by more than two variables, a partial correlation coefficient measures the strength of dependence between a pair of variables that is not accounted for by the way in which they both change in response to variations in a selected subset of the other variables.

Decorrelation

It is always possible to remove the correlation between random variables with a linear transformation, even if the relationship between the variables is nonlinear. A presentation of this result for population distributions is given by Cox & Hinkley.[28]

A corresponding result exists for sample correlations, in which the sample correlation is reduced to zero. Suppose a vector of *n* random variables is sampled *m* times. Let *X* be a matrix where $X_{i,j}$ is the *j*th variable of sample *i*. Let $Z_{m,m}$ be an *m* by *m* square matrix with every element 1. Then *D* is the data transformed so every random variable has zero mean, and *T* is the data transformed so all variables have zero mean and zero correlation with all other variables – the sample covariance matrix of *T* will be the identity matrix. This has to be further divided by the standard deviation to get unit variance. The transformed variables will be uncorrelated, even though they may not be independent.

$$D = X - \frac{1}{m} Z_{m,m} X$$

$$T = D(D^T D)^{-\frac{1}{2}},$$

where an exponent of −1/2 represents the matrix square root of the inverse of a matrix. The covariance matrix of T will be the identity matrix. If a new data sample x is a row vector of n elements, then the same transform can be applied to x to get the transformed vectors d and t:

$$d = x - \frac{1}{m} Z_{1,m} X,$$

$$t = d(D^T D)^{-\frac{1}{2}}.$$

This decorrelation is related to principal components analysis for multivariate data.

See also

- Association (statistics)
- Correlation and dependence
- Disattenuation
- Distance correlation
- Maximal information coefficient
- Multiple correlation
- Normally distributed and uncorrelated does not imply independent
- Partial correlation
- Quadrant count ratio
- RV coefficient
- Spearman's rank correlation coefficient

References

1. " "SPSS Tutorials: Pearson Correlation" " (http://libguides.library.kent.edu/SPSS/PearsonCorr). Retrieved 2017-05-14.
2. See:
   - As early as 1877, Galton was using the term "reversion" and the symbol "*r*" for what would become "regression". F. Galton (5, 12, 19 April 1877) "Typical laws of heredity," *Nature*, **15** (388, 389, 390) : 492–495 ; 512–514 ; 532–533. In the "Appendix" on page 532 (https://books.google.com/books?id=eskKAAAAYAAJ&pg=PA512#v=onepage&q&f=false), Galton uses the term "reversion" and the symbol *r*.
   - (F. Galton) (24 September 1885), "The British Association: Section II, Anthropology: Opening address by Francis Galton, F.R.S., etc., President of the Anthropological Institute, President of the Section," (https://books.google.com/books?id=lN3RjXLUuWsC&pg=PA499#v=onepage&q&f=false) *Nature*, **32** (830) : 507–510.
   - Galton, F. (1886) "Regression towards mediocrity in hereditary stature," (https://books.google.com/books?id=JPcRAAAAYAAJ&pg=PA246#v=onepage&q&f=false) *Journal of the Anthropological Institute of Great Britain and Ireland*, **15** : 246–263.
3. Karl Pearson (20 June 1895) "Notes on regression and inheritance in the case of two parents," (https://books.google.com/books?id=60aL0zlT-90C&pg=PA240#v=onepage&q&f=false) *Proceedings of the Royal Society of London*, **58** : 240–242.

4. Stigler, Stephen M. (1989). "Francis Galton's Account of the Invention of Correlation". *Statistical Science*. **4** (2): 73–79. JSTOR 2245329 (https://www.jstor.org/stable/2245329). doi:10.1214/ss/1177012580 (https://doi.org/10.1214%2Fss%2F1177012580).
5. Real Statistics Using Excel: Correlation: Basic Concepts (http://www.real-statistics.com/correlation/basic-concepts-correlation/), retrieved 2015-02-22
6. Moriya, N. (2008). Fengshan Yang, ed. "Noise-Related Multivariate Optimal Joint-Analysis in Longitudinal Stochastic Processes". Nova Science Publishers, Inc.: 223–260. ISBN 978-1-60021-976-4.
7. Garren, Steven T (15 June 1998). "Maximum likelihood estimation of the correlation coefficient in a bivariate normal model with missing data" (http://www.sciencedirect.com/science/article/pii/S0167715298000352). *Statistics & Probability Letters*. Elsevier. **38** (3): 281–288. doi:10.1016/S0167-7152(98)00035-2 (https://doi.org/10.1016%2FS0167-7152%2898%2900035-2). Retrieved 5 December 2015.
8. Schmid Jr., John (December 1947). "The Relationship between the Coefficient of Correlation and the Angle Included between Regression Lines". *The Journal of Educational Research*. **41** (4). JSTOR 27528906 (https://www.jstor.org/stable/27528906).
9. Rummel, R. J. (1976). "Understanding Correlation" (http://www.hawaii.edu/powerkills/UC.HTM).
10. Buda, Andrzej; Jarynowski, Andrzej (December 2010). *Life time of correlations and its applications*. Wydawnictwo Niezależne. pp. 5–21. ISBN 9788391527290.
11. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.)
12. Rahman, N. A. (1968) *A Course in Theoretical Statistics*, Charles Griffin and Company, 1968
13. Kendall, M. G., Stuart, A. (1973) *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*, Griffin. ISBN 0-85264-215-6 (Section 31.19)
14. Soper, H. E.; Young, A. W.; Cave, B. M.; Lee, A.; Pearson, K. (1917). "On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A co-operative study". *Biometrika*. **11**: 328–413. doi:10.1093/biomet/11.4.328 (https://doi.org/10.1093%2Fbiomet%2F11.4.328).
15. Kenney, J. F. and Keeping, E. S., *Mathematics of Statistics*, Pt. 2, 2nd ed. Princeton, NJ: Van Nostrand, 1951.
16. Correlation Coefficient – Bivariate Normal Distribution (http://mathworld.wolfram.com/CorrelationCoefficientBivariateNormalDistribution.html)
17. Wilcox, Rand R. (2005). *Introduction to robust estimation and hypothesis testing*. Academic Press.
18. Devlin, Susan J; Gnanadesikan, R; Kettenring J.R. (1975). "Robust Estimation and Outlier Detection with Correlation Coefficients". *Biometrika*. **62** (3): 531–545. JSTOR 2335508 (https://www.jstor.org/stable/2335508). doi:10.1093/biomet/62.3.531 (https://doi.org/10.1093%2Fbiomet%2F62.3.531).
19. Huber, Peter. J. (2004). *Robust Statistics*. Wiley.
20. Katz., Mitchell H. (2006) *Multivariable Analysis – A Practical Guide for Clinicians*. 2nd Edition. Cambridge University Press. ISBN 978-0-521-54985-1. ISBN 0-521-54985-X doi:10.2277/052154985X (https://dx.doi.org/10.2277%2F052154985X)
21. Hotelling, H. (1953). "New Light on the Correlation Coefficient and its Transforms". *Journal of the Royal Statistical Society. Series B (Methodological)*. **15** (2): 193–232. JSTOR 2983768 (https://www.jstor.org/stable/2983768).
22. Olkin, Ingram; Pratt,John W. (March 1958). "Unbiased Estimation of Certain Correlation Coefficients". *The Annals of Mathematical Statistics*. **29** (1): 201–211. JSTOR 2237306 (https://www.jstor.org/stable/2237306). doi:10.1214/aoms/1177706717 (https://doi.org/10.1214%2Faoms%2F1177706717)..
23. http://sci.tech-archive.net/Archive/sci.stat.math/2006-02/msg00171.html
24. A MATLAB Toolbox for computing Weighted Correlation Coefficients (http://www.mathworks.com/matlabcentral/fileexchange/20846)
25. Nikolić, D; Muresan, RC; Feng, W; Singer, W (2012). "Scaled correlation analysis: a better way to compute a cross-correlogram" (http://www.danko-nikolic.com/wp-content/uploads/2012/03/Scaled-correlation-analysis.pdf) (PDF). *European Journal of Neuroscience*: 1–21. doi:10.1111/j.1460-9568.2011.07987.x (https://doi.org/10.1111%2Fj.1460-9568.2011.07987.x).
26. Fulekar (Ed.), M.H. (2009) *Bioinformatics: Applications in Life and Environmental Sciences*, Springer (pp. 110) ISBN 1-4020-8879-5
27. Jammalamadaka, S. Rao; SenGupta, A. (2001). *Topics in circular statistics* (https://books.google.com/books?id=sKqWMGqQXQkC&printsec=frontcover&dq=Jammalamadaka+Topics+in+circular&hl=en&ei=iJ3QTe77NKL00gGdyqHoDQ&sa=X&oi=book_result&ct=result&resnum=1&ved=0CDcQ6AEwAA#v=onepage&q&f=false). New Jersey: World Scientific. p. 176. ISBN 981-02-3778-2. Retrieved 2016-09-21.
28. Cox, D.R., Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman & Hall (Appendix 3) ISBN 0-412-12420-3

External links

- cocor – A free web interface and R package for the statistical comparison of two dependent or independent correlations with overlapping or non-overlapping variables (http://comparingcorrelations.org)
- Interactive Flash simulation on the correlation of two normally distributed variables. (http://nagysandor.eu/AsimovTeka/correlation_en/index.html)
- Correlation coefficient calculator – linear regression (http://www.hackmath.net/en/calculator/linear-regression)
- "Critical values for Pearson's correlation coefficient (large table)" (http://frank.mtsu.edu/~dkfuller/tables/correlationtable.pdf) (PDF).
- Guess the Correlation – A game where players guess how correlated two variables in a scatter plot are, in order to gain a better understanding of the concept of correlation. (http://guessthecorrelation.com)

Wikiversity has learning resources about ***Linear correlation***

Categories: Covariance and correlation | Parametric statistics | Statistical ratios

---