

Covariance and Correlation

Understanding how two variables vary together

Ratiomera Statistics

Table of contents

Course-Page Summary	2
From Paired Deviations to Correlation	2
What to Check Before Interpreting r	2
The Bridge to Regression	3
Expanded Reference	4
Purpose and foundations	4
Core ideas	4
Formula guide	5
Reading the explanatory figure	6
Interpretation checklist	7
How this topic connects	7

Course-Page Summary

Covariance and correlation describe how two quantitative variables vary together across correctly paired cases. Always inspect the scatterplot first. A coefficient is a summary of that plot, not a replacement for it.

From Paired Deviations to Correlation

For paired observations (x_i, y_i) , sample covariance is

$$s_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

Products are positive when both deviations point in the same direction and negative when they point in opposite directions. The sign therefore records direction. Its size depends on the measurement units, which makes covariances difficult to compare across scales.

Pearson's correlation standardizes the covariance:

$$r = \frac{s_{XY}}{s_X s_Y}.$$

It lies between -1 and $+1$. The sign gives the direction of the linear pattern, and $|r|$ describes how closely the points follow a straight line. A value near zero means little linear association; a strong curved pattern can still have a small Pearson correlation.

What to Check Before Interpreting r

Table 1: Read the data pattern and study design before reading the coefficient.

Check	Why it matters
Paired rows	Each x_i must stay with its own y_i
Scatterplot shape	Pearson's r summarizes linear association, not every form of dependence
Unusual and influential points	One point can change the coefficient substantially
Observed range	Restricting either variable's range can weaken the observed coefficient
Subgroups	A combined pattern can hide different patterns within groups
Measurement and sampling	Error, selection, and poor coverage can distort the relationship

Spearman's rank correlation applies Pearson's idea to ranks. It describes a monotonic pattern, meaning that one variable tends to move consistently upward or downward as the other changes, without requiring a straight line. Tied values receive appropriate shared ranks.

For inference about a population Pearson correlation under the model used here,

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

The test asks whether the population linear correlation could be zero under the stated assumptions. It does not turn an observational association into a causal effect.

The Bridge to Regression

Correlation is symmetric and unit-free. Simple regression keeps the same paired pattern but gives the variables different roles: one becomes the predictor and the other the outcome. Topic 5 turns the point cloud into a fitted line, restores original units, and studies the residuals that show what the line misses.

Expanded Reference

Purpose and foundations

Covariance and correlation describe how two numerical variables vary together across the same cases. Each case must contribute a matched pair (x_i, y_i) . A scatterplot is the essential starting point: place one variable on the horizontal axis, the other on the vertical axis, and represent every case by one point. The resulting cloud shows direction, form, strength, clusters, and unusual observations in a way that a single coefficient cannot.

A positive association means that cases with larger x values tend to have larger y values. A negative association means that larger x values tend to accompany smaller y values. A value near zero for a linear coefficient means little linear association, not necessarily no relationship. Curved patterns, separate subgroups, or a restricted range can all make a coefficient incomplete or misleading.

Covariance begins with deviations from the two means. A case contributes positively when both values are above their means or both are below. It contributes negatively when one value is above its mean and the other is below. Averaging these cross-products gives the sample covariance. Its sign is informative, but its magnitude depends on the measurement units. Measuring hours in minutes, for example, changes the covariance even though the underlying pairing of cases has not changed.

Feature	Covariance	Pearson correlation
Direction	Sign shows positive or negative co-variation	Sign shows positive or negative linear association
Scale	Depends on both variables' units	Unit-free because both variables are standardized
Numerical range	Not restricted to a fixed interval	Always between -1 and 1
Main role	Building block for association and regression	Comparable summary of linear direction and strength

Core ideas

Pearson's correlation r standardizes covariance by dividing by the two sample standard deviations. A value near 1 indicates that the points follow a strong positive straight-line pattern. A value near -1 indicates a strong negative straight-line pattern. A value near zero indicates little straight-line pattern. The coefficient describes the sample. The population correlation is commonly written ρ , and inference is needed when the goal is a population conclusion.

Spearman's rank correlation replaces observed values with their ranks and assesses whether the variables follow a **monotonic** relationship. Monotonic means that the tendency moves consistently in one direction: as one variable increases, the other generally increases, or it generally decreases. The pattern may curve while preserving that order. Spearman correlation can therefore remain high for a curved monotonic association that Pearson correlation summarizes less fully. Neither coefficient represents a U-shaped relationship well because the direction reverses across the range.

Diagnostic question	What to inspect	Why it changes interpretation
Is the form approximately linear?	Scatterplot and possible smooth pattern	Pearson r summarizes a straight-line tendency
Are unusual points influential?	Labeled scatterplot and sensitivity comparison	One distant point can change direction or magnitude
Are groups mixed together?	Colors or panels for meaningful groups	A pooled association may differ from within-group associations

Diagnostic question	What to inspect	Why it changes interpretation
Is the observed range restricted?	Variable ranges and sampling process	Limited variation can weaken an observed coefficient
Is the pairing valid?	Case identifiers and measurement timing	Correlation requires both values to refer to the same case

Correlation does not establish causation. An observed association may reflect a direct influence, a reverse direction of influence, a third variable related to both, selection into the sample, measurement artifacts, or chance. Temporal order and a credible research design provide information that a coefficient alone cannot. Even when causal reasoning is not intended, the substantive context still determines whether the paired variables and their interpretation make sense.

When a population conclusion is intended, the sample coefficient can be tested against $H_0 : \rho = 0$. Keep that test separate from magnitude and practical meaning. A small p-value concerns compatibility with a zero population correlation under the model; it does not make the association large, important, or causal.

Formula guide

The sample covariance averages paired cross-products of deviations, using $n - 1$ in the denominator:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

The sign of each product records whether the two deviations point in the same or opposite directions. Large deviations receive more weight because their product has greater magnitude.

Pearson's sample correlation divides covariance by the product of the two sample standard deviations:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

When corrected sums are available, the same coefficient can be calculated directly as

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

Both forms require the same paired cases in the two variables. Reordering one column without its partner destroys the pairs and changes the question.

The same calculation can be written as a sum of products of standardized scores. This makes the unit-free nature of the coefficient visible:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n z_{xi} z_{yi}$$

For a sample without tied ranks, Spearman's rank correlation can be calculated from the difference d_i between each case's two ranks:

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

With tied ranks, calculate Pearson correlation on the assigned ranks instead. In all forms, the coefficient must be read with the scatterplot or rank pattern that produced it.

To test the population Pearson correlation against $H_0 : \rho = 0$, use

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

The alternative determines whether the reference area is one-sided or two-sided. The calculation relies on independent paired cases, a relationship for which a linear Pearson summary is defensible, and the absence of design or influential-point problems that would invalidate the interpretation.

Result to report	Question it answers	What it cannot establish alone
Scatterplot form	What pattern, clusters, range, and unusual points are visible?	A population conclusion
r or r_s	What direction and linear or monotonic strength appears in the sample?	Causation
t , df , and p-value	How compatible is the sample Pearson coefficient with $\rho = 0$ under the model?	Practical importance or a large association

Reading the explanatory figure

A Scatterplot Reveals What One Coefficient Can Miss

Spearman follows monotonic order; neither coefficient summarizes a U-shape well

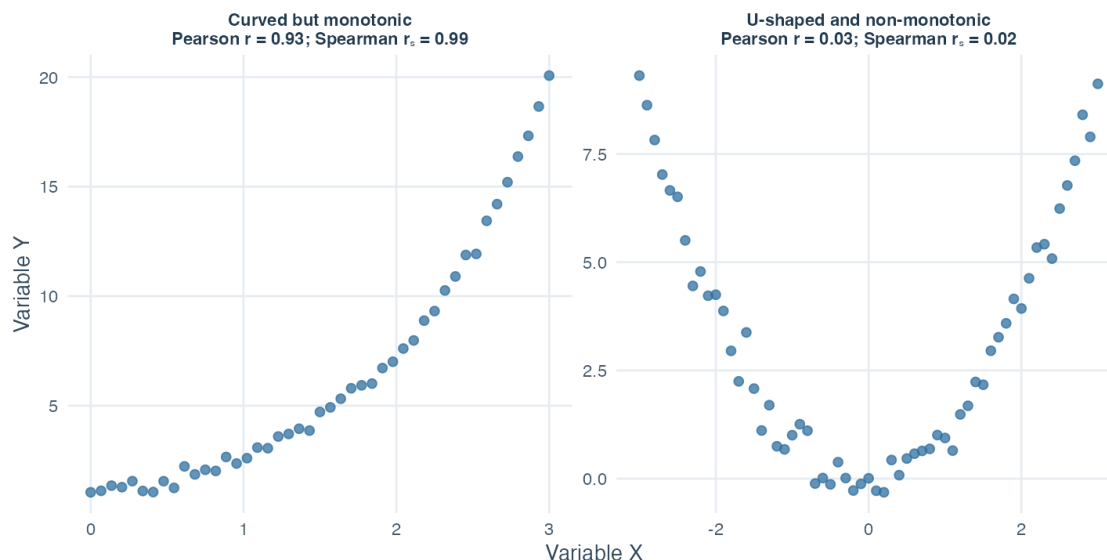


Figure 1: Two scatterplots compare a curved monotonic pattern with a U-shaped pattern and report Pearson and Spearman correlations above each panel.

The left panel rises throughout the observed range. The increase is curved rather than straight, but the ordering is highly consistent: larger x values nearly always accompany larger y values. Spearman's correlation is therefore close to one because ranks preserve this upward order. Pearson's correlation is also strongly positive, but its focus remains the overall straight-line component of the pattern. The visible curve tells you that a straight-line summary does not capture every feature.

The right panel is U-shaped. Moving from the far left toward the center, y decreases as x increases. Moving from the center toward the far right, y increases as x increases. These opposite directions cancel in both Pearson and Spearman calculations, producing values near zero. Yet the variables have a pronounced relationship. The correct conclusion is not “no association.” It is that neither a linear nor a monotonic coefficient summarizes this form well.

The connecting lines in the figure help reveal the order of the points; they are not fitted regression lines. The coefficients printed above the panels describe the displayed simulated values. They are teaching results, not estimates from real participants. This example is a reminder to let the plot identify the form and then let a coefficient summarize the feature it was designed to measure.

Interpretation checklist

Confirm that the variables are numerical or that a rank-based analysis is appropriate. Verify that the values are paired by case. Inspect a scatterplot before calculating a coefficient. Describe direction, form, strength, clusters, range, and unusual points. Choose Pearson correlation for a substantively meaningful linear summary and Spearman correlation for a monotonic rank-based summary. Report the sample size and coefficient, and add an interval or test when a population conclusion is required.

Do not label fixed universal cutoffs as weak, moderate, or strong without context. A coefficient's practical meaning depends on measurement reliability, field, design, and consequences. Check how the result changes when an influential point or meaningful subgroup is examined, but do not remove cases merely to improve a coefficient. Avoid causal verbs unless the design supports them.

How this topic connects

Covariance is the bridge from descriptive variability to regression. In simple linear regression, the slope can be written as covariance between predictor and outcome divided by predictor variance. Correlation standardizes the same paired tendency, while regression keeps the outcome's unit and gives a fitted change in that outcome for a one-unit predictor change.

Partial correlation later asks how two variables remain associated after linear adjustment for a third. Multiple regression extends the same logic by estimating the conditional association of each predictor while holding the others fixed. Analysis of variance also belongs to this family: it explains variation in an outcome using group membership rather than beginning with a numerical predictor. The shared question is how outcome variation aligns with information supplied by one or more predictors.

Document ID: topic-04-covariance-correlation-summary-en

Notes 1 of 3

Covariance and Correlation | Topic 4

Notes 2 of 3

Covariance and Correlation | Topic 4

Notes 3 of 3

Covariance and Correlation | Topic 4