

What is Restriction of Range?

Authored by
stats writer

December 17, 2025

RECOMMENDED CITATION

stats writer (2025). *What is Restriction of Range?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=107654>

The statistical phenomenon known as Restriction of Range is a crucial concern in quantitative analysis, specifically describing situations where the observed variability, or spread, of scores for a variable is artificially curtailed relative to the true population distribution. This limitation fundamentally compromises the integrity of bivariate statistics, particularly the calculation of the correlation coefficient. When range restriction occurs, the resulting correlation measure typically underestimates the true strength of the relationship that exists in the wider population, leading to potentially incorrect conclusions about predictive validity or causal links.

Often in statistical research, we are fundamentally interested in measuring the degree of association, or correlation, between two variables. This measurement is vital for understanding complex phenomena and making reliable predictions. Correlation analysis primarily helps us determine the following key attributes of a relationship:

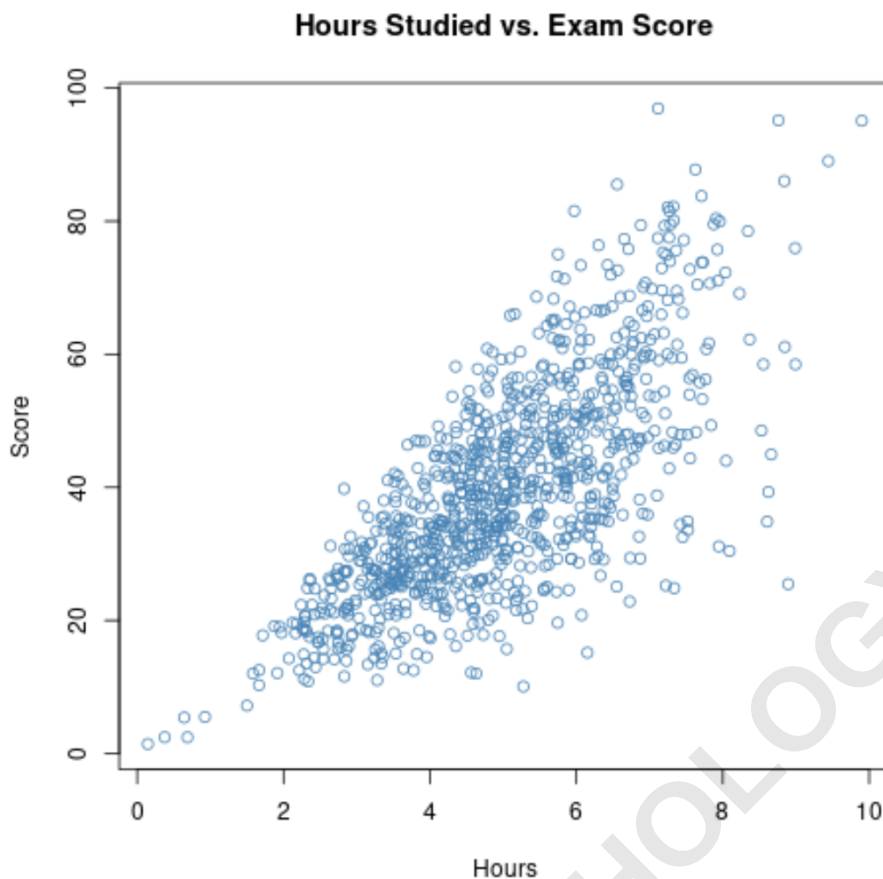
The **direction** of the relationship between two variables. This addresses the question: *As one variable increases, does the other variable tend to increase (positive correlation) or decrease (negative correlation)?*

The **strength** of the relationship between two variables. This quantifies *How closely do the two variables change in value?* A high strength indicates a reliable and tight predictive link between them.

Unfortunately, one pervasive methodological problem that can drastically distort the measured correlation is known as **restriction of range** (ROR). This statistical bias occurs when the *range* of values measured for one or both variables in the observed sample is significantly limited compared to the true range present in the full population, typically due to selection procedures or sampling errors.

For example, suppose we intend to measure the correlation between *hours studied* (Variable X) and *exam score* (Variable Y) for students across an entire academic institution.

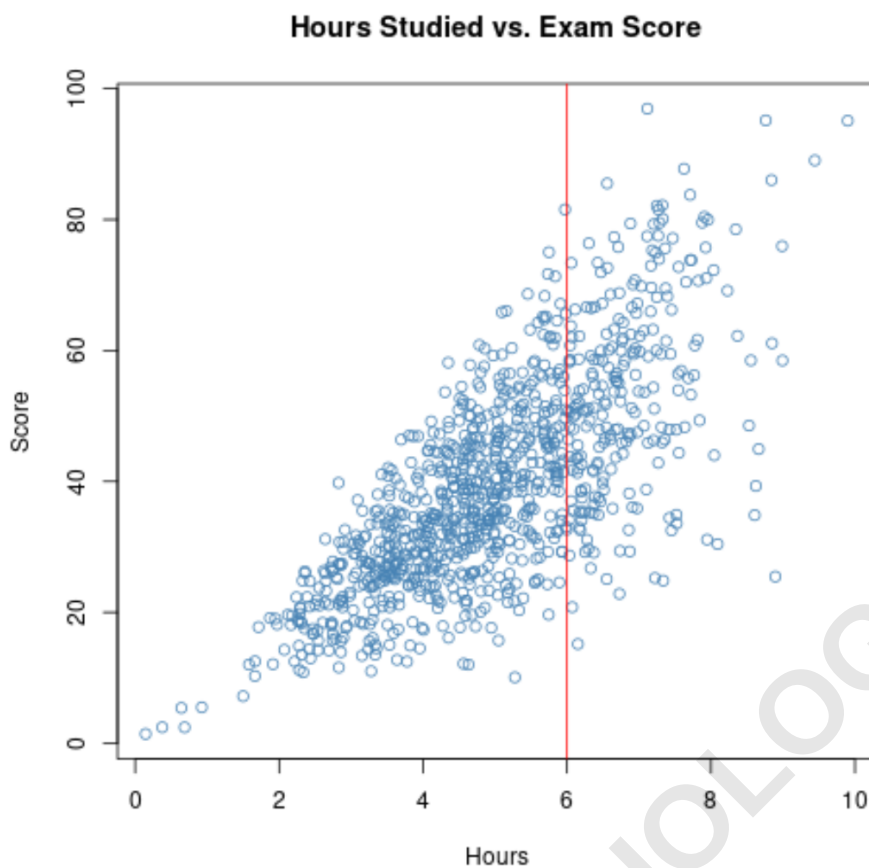
If we gather comprehensive data on these variables for all 1,000 students in the school, representing the full spectrum of academic effort, we might find that the population correlation between *hours studied* and *exam score* is a robust **0.73**. This figure represents the true, strong positive relationship across the entire population.



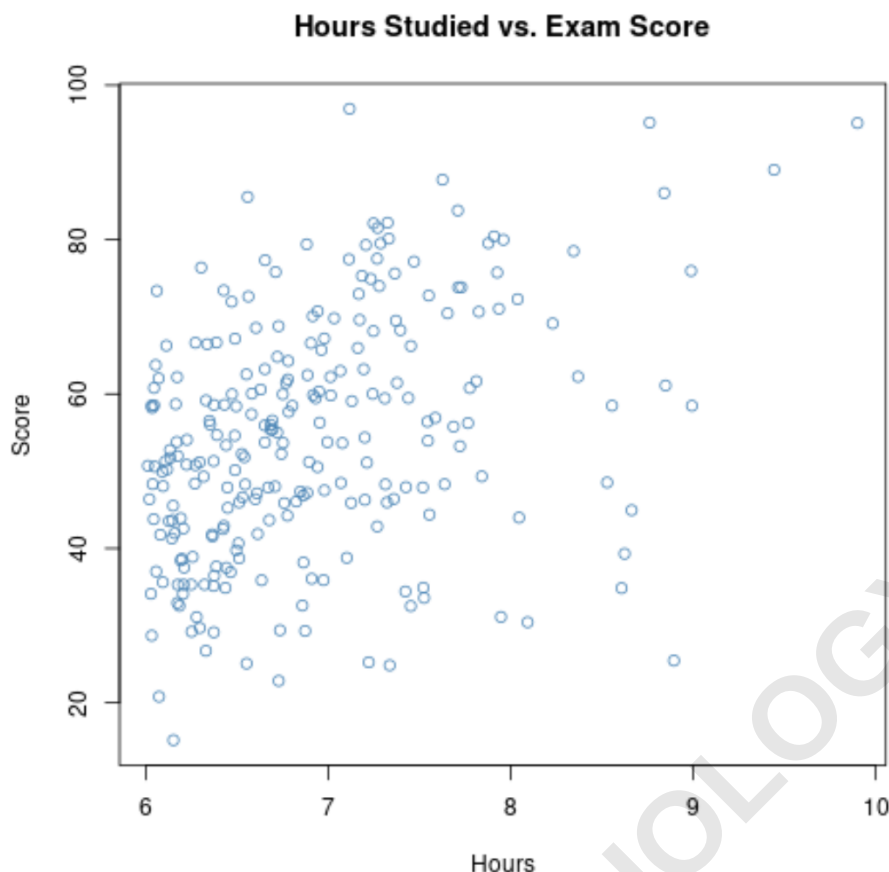
This initial correlation is quite high, signifying a strong positive relationship. Visually, the scatterplot shows a clear, tight linear trend, confirming that as students increase their study time, they reliably achieve higher exam scores.

However, consider the impact of sampling bias. Suppose we only collect data for students enrolled in highly competitive Honors courses. It is highly probable that, due to selection effects, every student in this cohort studied for at least 6 hours, effectively truncating the lower end of the "hours studied" variable.

Consequently, if we calculate the correlation using only data from these selected students, we are operating within a severely restricted range for the variable *hours studied*. The sample fails to capture the high degree of variance that exists in the general student body.



If we specifically isolate and analyze the scatterplot data confined to the range where *Hours Studied* is greater than 6, the resulting plot illustrates the compression of data points:



The correlation calculated based on this restricted subset drops dramatically to **0.37**, which is significantly lower than the true population correlation of **0.73**.

Thus, relying solely on data gathered from the Honors students would lead to the false assumption that there is only a weak or moderate relationship between study effort and exam performance.

This result highlights a critical form of statistical bias because the true predictive validity was masked by the use of a **restricted range** for one of the key variables.

Real-World Examples of Restricted Range

The dilemma of a restricted range permeates many areas of applied research. This problem often arises in organizational or educational contexts where selection procedures naturally filter the pool of observable subjects. Here are detailed examples showcasing how ROR affects correlation estimates in practice:

1. Studies of Elite Physical Performance. If researchers are studying the relationship between a novel workout regimen and muscle mass gains, restricting the sample to only high-performance athletes presents a classic ROR scenario. These athletes likely already possess maximal or near-

maximal muscle mass and are highly motivated. Consequently, there will be a narrow range of muscle mass values available, and limited variance for the outcome variable. This narrow range severely limits the ability to calculate a meaningful correlation between the new regimen and the muscle mass produced, often leading to an underestimation of the regimen's effectiveness for a typical individual.

2. Evaluating Interventions for High-Performance Students. Researchers may be interested in assessing whether a certain advanced tutoring program positively affects standardized test scores. Students who opt into such programs are often already high-achieving, meaning they have little room for improvement in their scores. When researchers calculate the correlation between the hours spent in the tutoring program and the resulting increase in grades, the true correlation may be significantly understated. The restricted range of possible grade improvements limits the observable covariance, making the intervention appear less impactful than it might be for a wider population.

How to Account for Restricted Ranges

Given the serious nature of the statistical bias introduced by ROR, specific methods have been developed by psychometrician to statistically correct the observed correlation coefficient. These methods aim to estimate the correlation that would have been obtained had the sample been unrestricted.

One popular and reliable way to account for restricted ranges, particularly when selection is performed on only one variable and population parameters are known, is known as **Thorndike's Case 2**. This is a formula developed by the eminent psychometrician Robert L. Thorndike, designed to statistically "inflate" the restricted correlation back toward the true population value.

This formula provides an estimate of the true correlation between two variables (X and Y) and relies on the following calculation structure:

$$\text{True correlation} = \sqrt{(1 - (\text{SD}^2y_{\text{restricted}} - \text{SD}^2y_{\text{unrestricted}})) * (1 - r^2_{\text{restricted}})}$$

where the variables are defined as:

SD²y restricted: The squared standard deviation (variance) of the available data on the response variable y, obtained from the restricted sample.

SD²y unrestricted: The known squared standard deviation (variance) of the response variable for the entire, unrestricted population.

r²restricted: The squared correlation coefficient calculated from the available restricted data.

This formula has been rigorously tested and shown to be effective at producing unbiased estimates of the true correlation between two variables, provided the selection process only restricted one

variable and the population parameters are accurate.

Crucially, in order to successfully use this formula, researchers must possess an accurate estimate of the true population standard deviation (or variance) for the response variable. Without this external population metric, the correction cannot be accurately applied, potentially introducing new error or reinforcing the existing statistical bias.

ARABPSYCHOLOGY.COM