

# AP Statistics - Scatterplots and Correlation - Lesson

Common Core AP STAT 2.4, AP STAT 2.5, HSS-ID.B.6, HSS-ID.C.8, HSS-ID.C.9 · [maths-pdf.fr/en/](https://maths-pdf.fr/en/)

**A scatterplot displays two quantitative variables and is described by direction, form, strength, and unusual features in context. The correlation coefficient  $r$ , between -1 and 1, measures the strength and direction of a linear association only. Correlation has no units, is not resistant to outliers, and never proves that one variable causes the other.**

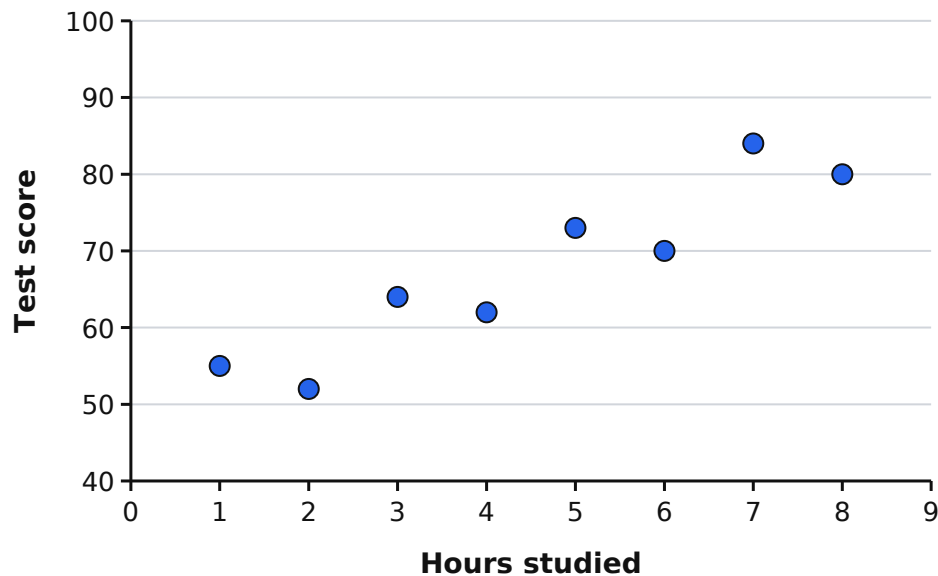
**AP Statistics** · Common Core State Standards: **AP STAT 2.4, AP STAT 2.5, HSS-ID.B.6, HSS-ID.C.8, HSS-ID.C.9**

## Key facts

- Explanatory variable on the x-axis, response variable on the y-axis.
- Describe direction, form, strength, and unusual features in context.
- $-1 \leq r \leq 1$ ;  $r$  near 0 means little linear association, not no association.
- $r$  is unitless, unchanged by swapping  $x$  and  $y$  or by linear changes of units, and not resistant.
- Correlation does not imply causation; look for lurking variables.

## What is a scatterplot?

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The **explanatory variable** goes on the horizontal axis and the **response variable** on the vertical axis. Each point is one individual.

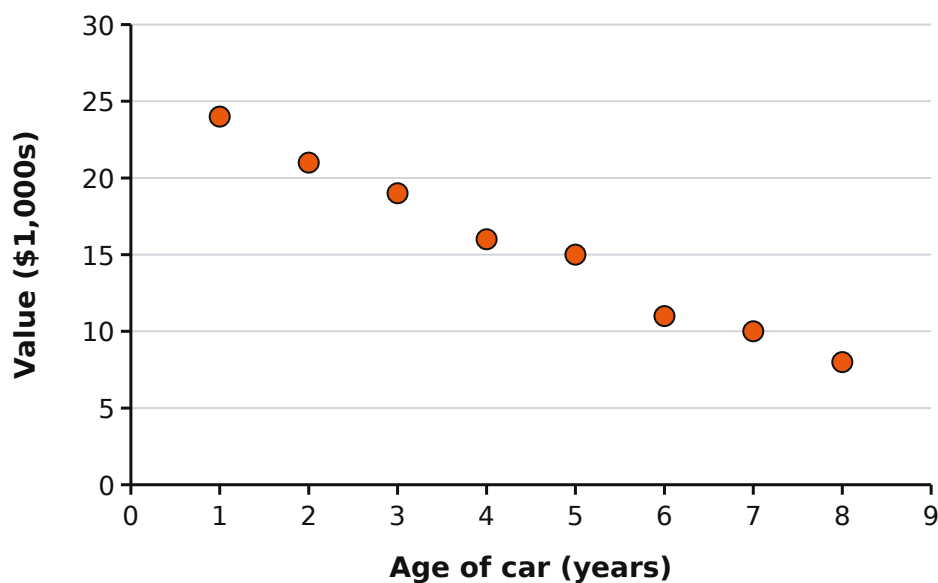


Positive, roughly linear, strong association between hours studied and test score ( $r = 0.934$ ).

### How do you describe a scatterplot?

Describe four things, in context: **direction** (positive or negative), **form** (linear or curved), **strength** (how closely the points follow the form), and **unusual features** such as outliers and clusters.

In the plot above, more hours studied go with higher scores, so the association is positive, roughly linear, and strong.



Negative, strong, nearly linear association: older cars are worth less ( $r = -0.994$ ).

## What is the correlation coefficient?

The **correlation**  $r$  measures the strength and direction of a *linear* association between two quantitative variables:

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

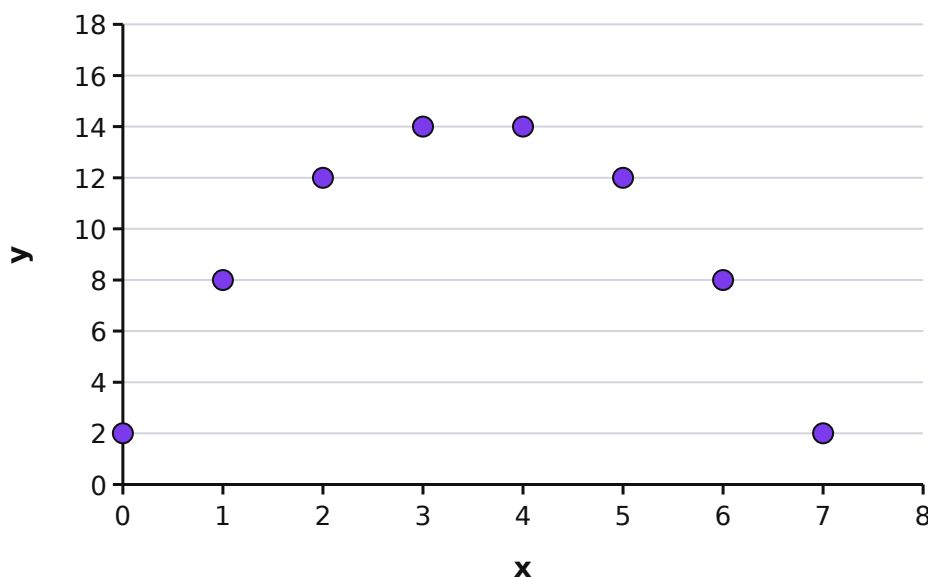
### Properties of $r$

$r$  is always between -1 and 1. The sign gives the direction.  $r$  has no units and does not change if you swap  $x$  and  $y$  or change units. It is not resistant to outliers, and it measures only linear association.

| $r$ near | Meaning                  |
|----------|--------------------------|
| 1        | strong positive linear   |
| 0.5      | moderate positive linear |
| 0        | no linear association    |
| -0.5     | moderate negative linear |
| -1       | strong negative linear   |

Rough guide to the strength of a linear association.

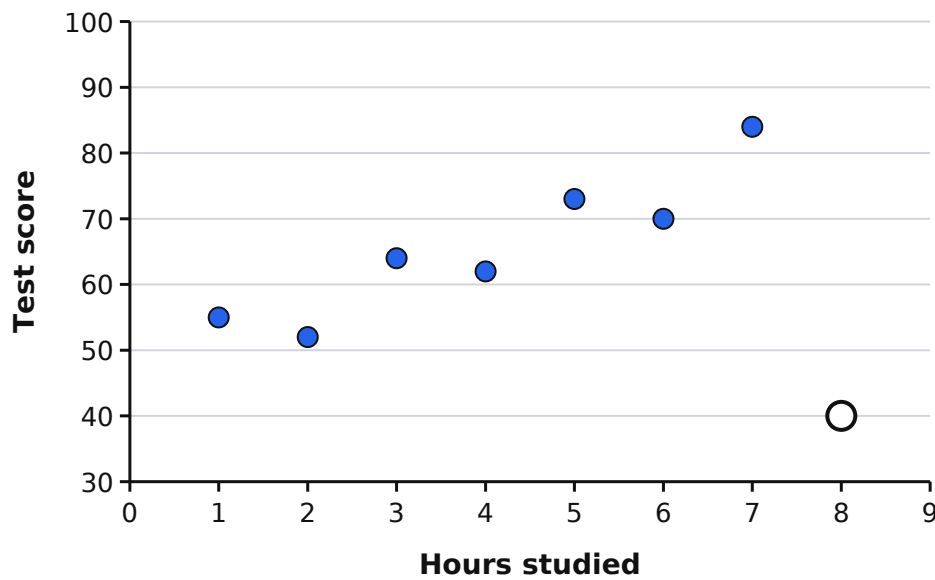
For the study-hours data,  $r = 0.934$ : a strong positive linear association. For the car-value data,  $r = -0.994$ : a strong negative linear association.



A clear but curved pattern. The correlation is near 0 ( $r = 0.000$ ), yet the variables are strongly related.

## What can go wrong with correlation?

Suppose one student's score is recorded as 40 instead of 80. The correlation drops sharply even though the other seven points still follow the trend. That is why you should look for unusual points and ask whether they are errors or real values. A single point can also hide a curved pattern or create a false impression of a strong line, so  $r$  is never a substitute for the graph.



One unusual point lowers the correlation from 0.934 to 0.179. Correlation is not resistant.

1. Always plot the data first:  $r$  cannot detect curved patterns.
2. Check for outliers, which can change  $r$  a lot.
3. Do not conclude causation from correlation. A lurking variable may drive both variables, like temperature affecting both ice cream sales and swimming accidents.

## Questions about scatterplots and correlation

### What does a correlation of 0 mean?

It means there is no linear association. The variables could still have a strong curved relationship, so always examine the scatterplot as well as  $r$ .

### Is a correlation of -0.9 weaker than +0.5?

No. The sign gives only the direction. The strength depends on the distance of  $r$  from 0, so -0.9 is much stronger than +0.5.

### Does a high correlation mean one variable causes the other?

No. Both variables may be influenced by a third, lurking variable, or the association may be coincidence. Only a well-designed randomized experiment can establish cause and effect.

### **Why does $r$ not change when I change units?**

$r$  is computed from z-scores, which are unitless. Converting inches to centimeters scales  $x$  and its standard deviation equally, leaving each z-score unchanged.

### **How do outliers affect correlation?**

An outlier can raise or lower  $r$  a lot, depending on whether it follows the trend. A point far from the pattern lowers  $r$ , while a point far out along the trend can raise it.