

## 14: Correlation

[Introduction](#) | [Scatter Plot](#) | [The Correlational Coefficient](#) | [Hypothesis Test](#) | [Assumptions](#) | [An Additional Example](#)

### Introduction

Correlation quantifies the extent to which two quantitative variables, X and Y, “go together.” When high values of X are associated with high values of Y, a positive correlation exists. When high values of X are associated with low values of Y, a negative correlation exists.

**Illustrative data set.** We use the data set `bicycle.sav` to illustrate correlational methods. In this cross-sectional data set, each observation represents a neighborhood. The X variable is socioeconomic status measured as the percentage of children in a neighborhood receiving free or reduced-fee lunches at school. The Y variable is bicycle helmet use measured as the percentage of bicycle riders in the neighborhood wearing helmets. Twelve neighborhoods are considered:

Neighborhood	X (% receiving reduced-fee lunch)	Y (% wearing bicycle helmets)
Fair Oaks	50	22.1
Strandwood	11	35.9
Walnut Acres	2	57.9
Discov. Bay	19	22.2
Belshaw	26	42.4
Kennedy	73	5.8
Cassell	81	3.6
Miner	51	21.4
Sedgewick	11	55.2
Sakamoto	2	33.3
Toyon	19	32.4
Lietz	25	38.4

There are twelve observations ( $n = 12$ ). Overall,  $\bar{x} = 30.83$  and  $\bar{y} = 30.883$ . We want to explore the relation between socioeconomic status and the use of bicycle helmets.

It should be noted that an outlier (84, 46.6) has been removed from this data set so that we may quantify the linear relation between X and Y.

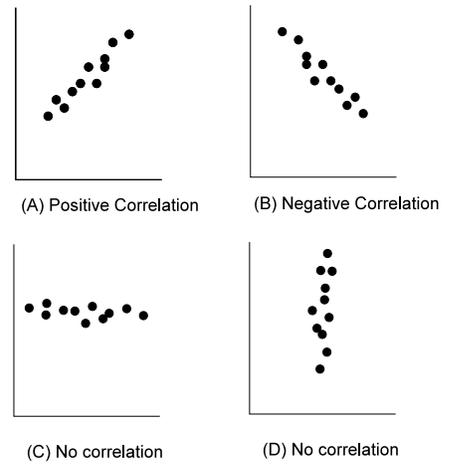
## Scatter Plot

The first step is create a scatter plot of the data. “There is no excuse for failing to plot and look.”<sup>1</sup>

In general, scatter plots may reveal a

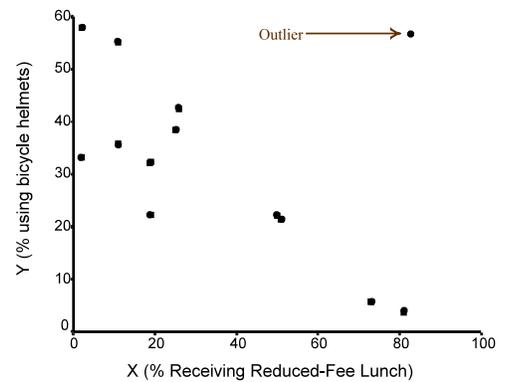
- **positive correlation** (high values of X associated with high values of Y)
- **negative correlation** (high values of X associated with low values of Y)
- **no correlation** (values of X are not at all predictive of values of Y).

These patterns are demonstrated in the figure to the right.



**Illustrative example.** A scatter plot of the illustrative data is shown to the right. The plot reveals that high values of X are associated with low values of Y. That is to say, as the number of children receiving reduced-fee meals at school increases, bicycle helmet use rates decrease’ a negative correlation exists.

In addition, there is an aberrant observation (“outlier”) in the upper-right quadrant. Outliers should *not* be ignored—it is important to say *something* about aberrant observations.<sup>2</sup> What should be said exactly depends on what can be learned and what is known. It is possible the lesson learned from the outlier is more important than the main object of the study. In the illustrative data, for instance, we have a low SES school with an envious safety record. What gives?



**Figure 2**

---

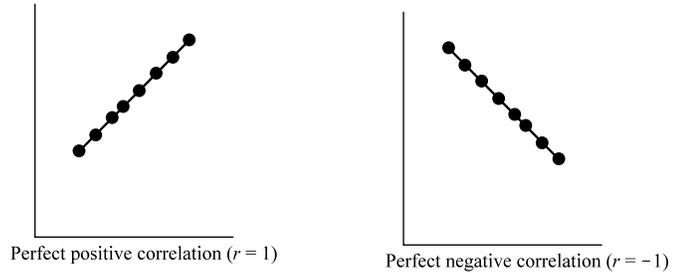
<sup>1</sup> Tukey, J. W. (1977). *EDA*. Reading, Mass.: Addison-Wesley, p. 43.

<sup>2</sup> Kruskal, W. H. (1959). Some Remarks on Wild Observations. <http://www.tufts.edu/~gdallal/out.htm>.

# Correlation Coefficient

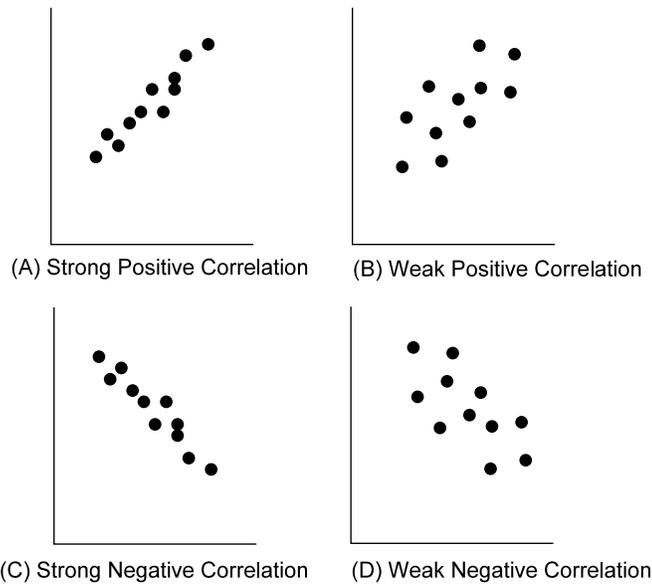
## The General Idea

Correlation coefficients (denoted  $r$ ) are statistics that quantify the relation between  $X$  and  $Y$  in unit-free terms. When all points of a scatter plot fall directly on a line with an upward incline,  $r = +1$ ; When all points fall directly on a downward incline,  $r = -1$ .



Such perfect correlation is seldom encountered. We still need to measure *correlational strength*, –defined as the *degree* to which data point adhere to an imaginary trend line passing through the “scatter cloud.” Strong correlations are associated with scatter clouds that adhere closely to the imaginary trend line. Weak correlations are associated with scatter clouds that adhere marginally to the trend line.

The closer  $r$  is to  $+1$ , the stronger the positive correlation. The closer  $r$  is to  $-1$ , the stronger the negative correlation. Examples of strong and weak correlations are shown below. **Note:** Correlational strength can *not* be quantified visually. It is too subjective and is easily influenced by axis-scaling. The eye is not a good judge of correlational strength.



## Pearson's Correlation Coefficient

To calculate a correlation coefficient, you normally need three different sums of squares (SS). The sum of squares for variable X, the sum of square for variable Y, and the sum of the cross-product of XY.

The **sum of squares for variable X** is:

$$SS_{XX} = \sum (x_i - \bar{x})^2 \quad (1)$$

This statistic keeps track of the spread of variable X. For the illustrative data,  $\bar{x} = 30.83$  and  $SS_{XX} = (50-30.83)^2 + (11-30.83)^2 + \dots + (25-30.83)^2 = 7855.67$ . Since this statistic is the numerator of the variance of X ( $s_x^2$ ), it can also be calculated as  $SS_{XX} = (s_x^2)(n-1)$ . Thus,  $SS_{XX} = (714.152)(12-1) = 7855.67$ .

The **sum of squares for variable Y** is:

$$SS_{YY} = \sum (y_i - \bar{y})^2 \quad (2)$$

This statistic keeps track of the spread of variable Y and is the numerator of the variance of Y ( $s_y^2$ ). For the illustrative data  $\bar{y} = 30.883$  and  $SS_{YY} = (22.1-30.883)^2 + (35.9-30.883)^2 + \dots + (38.4-30.883)^2 = 3159.68$ . An alternative way to calculate the sum of squares for variable Y is  $SS_{YY} = (s_y^2)(n-1)$ . Thus,  $SS_{YY} = (287.243)(12-1) = 3159.68$ .

Finally, the **sum of the cross-products ( $SS_{XY}$ )** is:

$$SS_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y}) \quad (3)$$

For the illustrative data,  $SS_{XY} = (50-30.83)(22.1-30.883) + (11-30.83)(35.9-30.883) + \dots + (25-30.83)(38.4-30.883) = -4231.1333$ . This statistic is analogous to the other sums of squares except that it is used to quantify the extent to which the two variables "go together".

The **correlation coefficient ( $r$ )** is

$$r = \frac{SS_{XY}}{\sqrt{(SS_{XX})(SS_{YY})}} \quad (4)$$

For the illustrative data,  $r = \frac{-4231.1333}{\sqrt{(7855.67)(3159.68)}} = -0.849$ .

## Interpretation of Pearson's Correlation Coefficient

The sign of the correlation coefficient determines whether the correlation is positive or negative. The magnitude of the correlation coefficient determines the strength of the correlation. Although there are no hard and fast rules for describing correlational strength, I [hesitatingly] offer these guidelines:

$0 <  r  < .3$	weak correlation
$.3 <  r  < .7$	moderate correlation
$ r  > 0.7$	strong correlation

For example,  $r = -0.849$  suggests a strong negative correlation.

**SPSS:** To calculate correlation coefficients click `Analyze > Correlate > Bivariate`. Then select variables for analysis. Several bivariate correlation coefficients can be calculated simultaneously and displayed as a correlation matrix. Clicking the Options button and checking "Cross-product deviations and covariances" computes sums of squares (Formulas 17.1 - 17.3).

## Coefficient of Determination

The coefficient of determination is the square of the correlation coefficient ( $r^2$ ). For illustrative data,  $r^2 = (-0.849)^2 = 0.72$ . This statistic quantifies the proportion of the variance of one variable "explained" (in a statistical sense, not a causal sense) by the other. The illustrative coefficient of determination of 0.72 suggests 72% of the variability in helmet use is explained by socioeconomic status.

## Hypothesis Test

The sample correlation coefficient  $r$  is the estimator of population correlation coefficient ( $\rho$ ). Recall that relations in samples do not necessarily depict the same in the population. For example, in Figure 6, the population of all dots demonstrates no correlation. If by chance the encircled points were sampled, an inverse association would appear. Thus, some samples cannot be relied on.

The null and alternative hypotheses are  $H_0: \rho = 0$  vs.  $H_1: \rho \neq 0$ .

If fixed-level testing is conducted, an  $\alpha$  level is selected.

The hypothesis can be tested with a  $t$  statistic:

$$t_{\text{stat}} = \frac{r}{se_r} \quad (5)$$

where  $se_r$  represents the **standard error the correlation coefficient**:

$$se_r = \sqrt{\frac{1-r^2}{n-2}} \quad (6)$$

Under the null hypothesis, this  $t$  statistic has  $n - 2$  degrees of freedom. Test results are converted to a  $p$  value before conclusions are drawn.

**Illustrative example.** For the illustrative data,  $se_r = \sqrt{\frac{1-(-0.849)^2}{12-2}} = 0.167$  and  $t_{\text{stat}} = \frac{-0.849}{0.167} = -5.08$ ,  $df = 12-2 = 10$ ,  $p = .00048$ . This provides evidence in support of the rejection of  $H_0$ .

**SPSS:** The  $p$  value — labeled “Sig. (2-tailed)” — is calculated as part of the Analyze > Correlate > Bivariate procedure and described on the prior page.

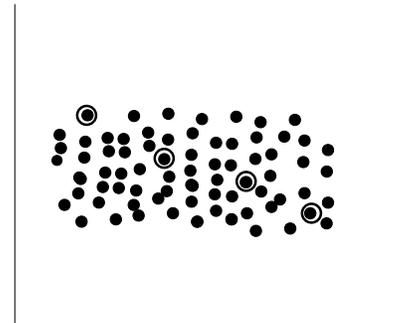


Figure 5

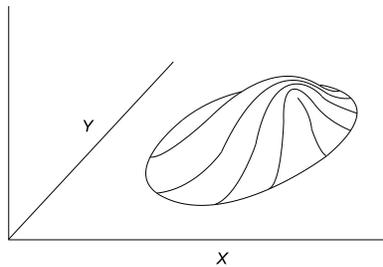
# Assumptions

We have in the past considered two types of assumptions:

- validity assumptions
- distributional assumptions

**Validity assumptions** require valid measurements, a good sample, unconfounded comparisons. These requirements are consistent throughout your studies.

The only **distributional assumption** needed to use  $r$  descriptively is linearity. Correlation coefficients assume the relation can be described with a straight line. The most direct way to determine linearity is to look at the scatterplot. Inferential methods require that the joint distribution of X and Y is **bivariate Normal**. This means the three-dimensional distribution of the scatter plot is bell-shaped from all angles:



Always remember that correlation does not equate with causation. The jump from a statistical relation to a causal relation requires a different framework.

## An Additional Example

The height of a child, of course, increases over time. Since the pattern of growth varies from child to child, one way to understand the general growth pattern is by using the average of childrens' heights by age. Data in the table to the right consider average height of participants in the Childhood Respiratory Disease study in East Boston, Massachusetts by age (fev.sav; data source = Rosner,1990, p. 39).

Data may be plotted to aid analysis. The figure at the bottom of the page shows average age with error bars based on the standard error of each mean (recall that  $se_{\text{mean}} = s / \sqrt{n}$ , and that this is a measure of the means precision – means based on large samples are more precise than means based on small samples, according to the “square root law”). Can the data be described by a straight line? (Not entirely, but perhaps in the range 4 to 11. Might a curved line between the ages of 4 and 17 be more useful? Perhaps we should split the data between boys and girls?)

Height (inches)

Age (years)	Mean	N	Std. Deviation
3	48.750	2	3.8891
4	48.778	9	1.4814
5	51.536	28	2.2316
6	53.311	37	2.3403
7	55.250	54	2.6346
8	58.306	85	3.1567
9	60.612	94	2.9285
10	62.475	81	3.3624
11	64.450	90	3.1747
12	65.202	57	3.4909
13	66.172	43	3.5843
14	67.520	25	3.3213
15	66.368	19	3.2009
16	67.385	13	3.6180
17	68.938	8	3.2562
18	66.000	6	3.5637
19	67.833	3	3.6171
Total	61.144	654	5.7035

