The Linear Regression (r^2) Test

Correlation – "Correlation does not imply causation"

Whenever you make a scatterplot graph, you can produce a statistic called the Pearson Product-moment Correlation Coefficient (*r*). This statistic (*r*) is a measure of the ***correlation*** (also called linear dependence) between two variables X and Y. The statistic gives a value between +1 and −1. It is widely used in the sciences as a measure of the *strength* of linear dependence between two variables. For example, the correlation can indicate how much the Y variable depends on what the X variable is doing. Mathematician, Karl Pearson, developed the correlation in the 1880s. The correlation coefficient is sometimes called "Pearson's *r*." The closer *r* is to +1 or -1, the stronger the correlation is between the two variables. However,

correlation does not imply causation.

For example, there may be a strong negative correlation between the mean temperature of the Earth over the last 190 years and the number of pirates in the Caribbean—the fewer the pirates, the warmer the Earth (Fig. 1). This does not mean that the number of pirates is the *cause* of global warming and to solve the global warming crisis, all we have to do is increase the number of pirates in the Caribbean. The *p*-value in Figure 1 tells us the probability of getting the *r* value by accident if the null hypothesis (H_0) is true. H_0 for correlation is that r = 0.

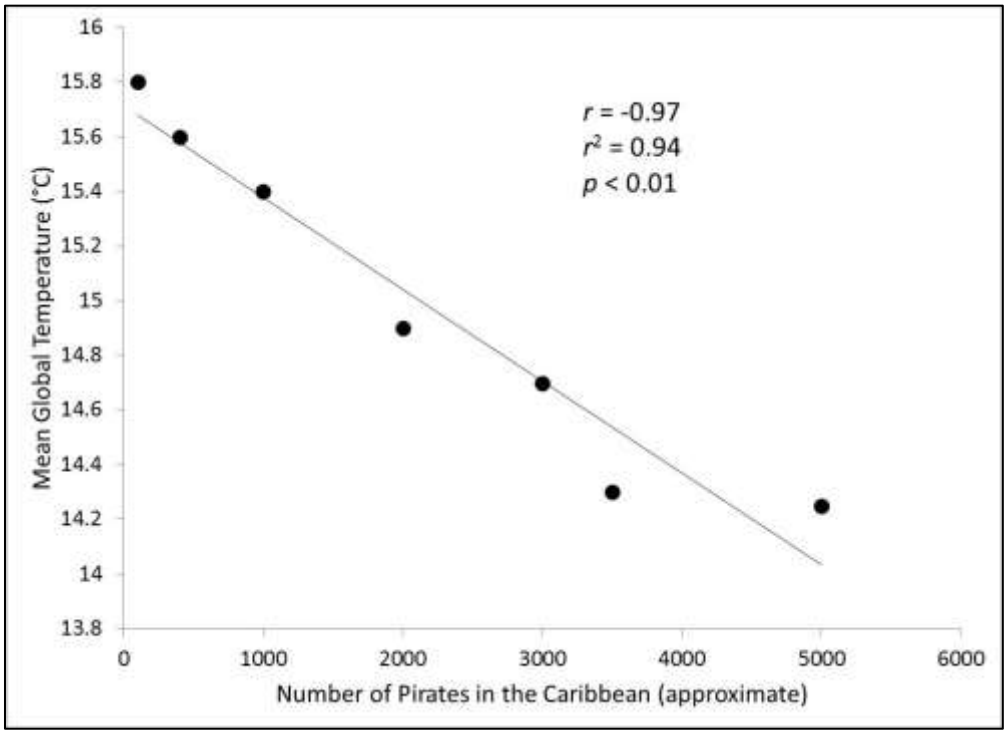

Figure 1. Mean global temperature (ºC) as a function of the approximate number of pirates in the Caribbean from 1820 to 2000. Line is linear regression. Statistics are Pearson's *r*, and coefficient of determination (r^2). Year 2000 data (100, 15.8 ºC) are to the upper left and historical data beginning in 1880 (5,000, 14.25 ºC) are to the lower right.

Linear Regression

Linear regression analysis uses the the coefficient of determination, the square of *r*, r^2. The r^2 value tells us the *strength* of the relationship between X and Y and can be interpreted in two equal ways:

- If your r^2 is 0.94, for example, then you can say the independent variable (X) predicts the dependent variable (Y) with 94% accuracy.
- You can also say that the pairs of coordinates have 94% of their variance in common. In other words, 94% of the variance in Y values is associated with the variance in X values.

Practice Scenario

Students noticed that some ponderosa pine trees (*Pinus ponderosa*) on a street had more ovulate cones (female pinecones) than other ponderosa pine trees. They hypothesized that the number of pine cones was a function of the age of the tree and predicted that taller (older) trees would have more cones than younger, shorter trees. To determine the height of a tree, they used the "old logger" method. A student held a stick the same length as the student's arm at a 90° angle to the arm and backed up until the tip of the stick "touched" the top of the tree. The distance the student was from the tree equaled the height of the tree. Using this method, the students measured the heights of 10 trees. Then, using binoculars, they counted the number of ovulate cones on each tree, recorded the data in Table 1, and performed all the calculations for determining *r*. ***You do not need to know how to perform these calculations.***

Table 1. Number of Ovulate Cones on Ponderosa Pine Trees of Different Heights

Tree No.	**Tree Height (m)**	**No. of Cones**	$\frac{x_i - \bar{x}}{s_x}$	$\frac{y_i - \bar{y}}{s_y}$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
1	10.5	75	1.226	1.034	1.267
2	7.2	68	0.133	0.790	0.105
3	4.3	59	−0.828	0.475	−0.393
4	7.9	46	0.364	0.021	0.008
5	3.8	8	−0.994	−1.307	1.298
6	8.3	56	0.497	0.370	0.184
7	3.4	25	−1.126	−0.713	0.802
8	4.1	13	−0.894	−1.132	1.012
9	12.3	15	1.822	−1.062	−1.934
10	6.2	89	−0.199	1.523	−0.303
Mean	$\bar{x}$ = 6.8	$\bar{y}$ = 45.4			$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ = 2.046
Standard deviation	s_x = 3.019	s_y = 28.625			

$$r = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1} = \frac{2.046}{10-1} = \frac{2.046}{9} = 0.227$$

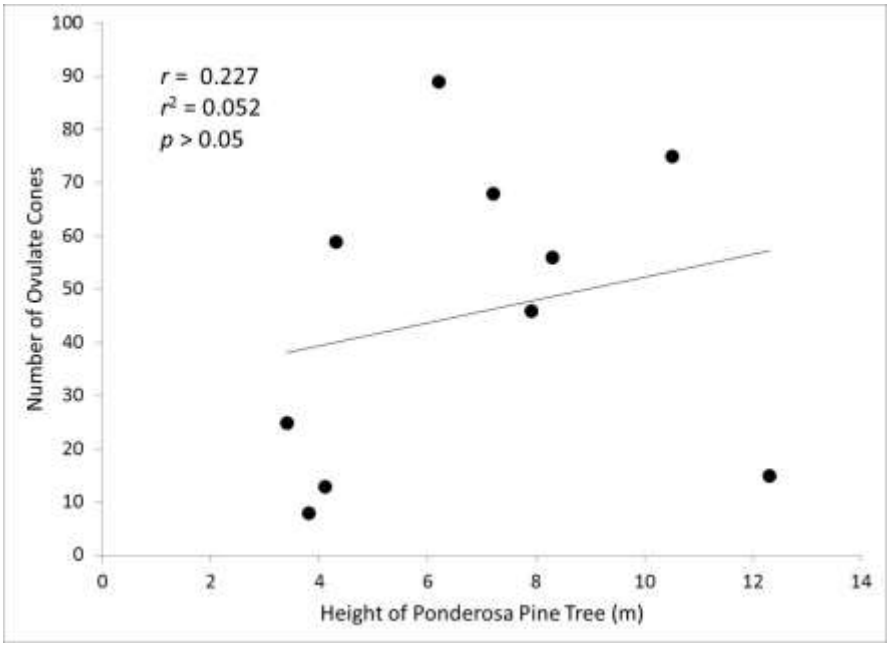

Figure 2: Number of ovulate cones on ponderosa pine trees as a function of tree height (m). Statistics are Pearson's correlation coefficient ($\boldsymbol{r}$) and the coefficient of determination ($\boldsymbol{r^2}$).

Just looking at the data points in Figure 2, it is hard to know whether there is a correlation or not. If there is a correlation, it is not very strong. Drawing the line of best fit suggests a positive correlation. This is clearly a case in which calculating r will help determine whether the correlation is statistically significant. In Table 3, r_{crit} is 0.632 for 8 degrees of freedom (10 – 2) at a probability level of 0.05. The calculated *r*-value is 0.227 which is smaller than the r_{crit} of 0.632, so the probability of getting a value of 0.227 purely by chance is greater than 0.05 ($p > 0.05$). Therefore, students cannot reject H_0 and can conclude that there is not a statistically significant association between the numbers of ovulate cones on ponderosa pine trees and the heights of the trees.

In EXCEL, you can create the scatterplot graph of the student's pine cone data. Insert a trend line for the data from the *Chart* option in the menu bar. Select *linear* for type of regression, and "display r^2 value" in *Options*.

More Practice! Graph and analyze the data in Table 2. In this scenario, students were curious to learn whether there is an association between amounts of algae in pond water and the water's clarity. They collected water samples from seven local ponds that seemed to differ in water clarity. To quantify the clarity of the water, they cut out a small disk from white poster board, divided the disk into four equal parts, and colored two of the opposite parts black; they then placed the disk in the bottom of a 100-milliliter graduated cylinder. For each sample, the students slowly poured pond water into the cylinder until the disk was no longer visible from above. In Table 2 they recorded the volume of water necessary to obscure the disk—the more water necessary to obscure the disk, the clearer the water. As a proxy for algae concentration, they extracted chlorophyll from the water samples and used a spectrophotometer to determine chlorophyll concentration.

What can the students conclude about the effect algae concentration seems to have on water clarity?

Table 2. Chlorophyll Concentration (μg/L) and Clarity of Pond Water

Pond	**Chlorophyll Concentration (x; μg/L)**	**Water Clarity (y; mL)**	$\frac{x_i - \bar{x}}{s_x}$	$\frac{y_i - \bar{y}}{s_y}$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
Sandy's	14	28	0.672	−0.656	−0.441
Herron	5	68	−0.956	1.077	−1.029
Tommy's	10	32	−0.052	−0.482	−0.025
Rocky	7	54	−0.594	0.470	−0.280
Fishing	17	18	1.214	−1.089	−1.323
Lost	16	25	1.033	−0.786	−0.812
Sunset	3	77	−1.318	1.467	−1.933
Mean	$\bar{x}$ = 10.29 μg/L	$\bar{y}$ = 43.14 mL			$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = -5.792$
Standard Deviation	s_x = 5.529	s_y = 23.083			

$$r = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1} = \frac{-5.792}{7-1} = \frac{-5.792}{6} = -0.965$$

Note: Water clarity is given as the volume of water in milliliters (mL) required to obscure a black-and-white disk at the bottom of a 100-milliliter graduated cylinder. A greater volume indicates clearer water.

Table 3. Critical values for *r* at three probability levels.

	Probability (*P*) that relationship occurred by chance/Probability the H_0 is true		
Degrees of Freedom (N pairs - 2)	**0.10**	**0.05**	**0.01**
1	0.988	0.997	0.999
2	0.900	0.950	0.990
3	0.805	0.878	0.959
4	0.729	0.811	0.917
5	0.669	0.754	0.874
6	0.662	0.707	0.834
7	0.582	0.666	0.798
8	0.549	0.632	0.765
9	0.521	0.602	0.735
10	0.497	0.576	0.708
11	0.476	0.553	0.684
12	0.458	0.532	0.661
13	0.441	0.514	0.641
14	0.426	0.497	0.623
15	0.412	0.482	0.606
16	0.400	0.468	0.590
17	0.389	0.456	0.575
18	0.378	0.444	0.561