CORRELATION COEFFICIENT

Data is a name for numerical information. You have learned how to make a scatter plot to display a set of data. Data analysis is a process of deriving information from data that helps you to answer the questions that interest you. The question that we are interested in is: Does temperature affect O-Rings? You have been studying about lines of best fit. There are many other measures of fit. The most commonly used measure is called the correlation coefficient, denoted by the letter $r$. The correlation coefficient is a measure of the linear association, or the way the data points cluster around the least squares regression line. In 1896, Karl Pearson gave a formula for calculating correlation coefficients and a justification for why his formula should give the best value, given that the data came from normal distributions.

\[
\text{slope} = \frac{n \left( \sum_{i=1}^{n} X_i Y_i \right) - \left( \sum_{i=1}^{n} X_i \right) \left( \sum_{i=1}^{n} Y_i \right)}{n \left( \sum_{i=1}^{n} X_i^2 \right) - \left( \sum_{i=1}^{n} X_i \right)^2}
\]

\[
y - \text{intercept} = \frac{\left( \sum_{i=1}^{n} X_i^2 \right) \left( \sum_{i=1}^{n} Y_i \right) - \left( \sum_{i=1}^{n} X_i \right) \left( \sum_{i=1}^{n} X_i Y_i \right)}{n \left( \sum_{i=1}^{n} X_i^2 \right) - \left( \sum_{i=1}^{n} X_i \right)^2}
\]

\[
r = \frac{n \left( \sum_{i=1}^{n} X_i Y_i \right) - \left( \sum_{i=1}^{n} X_i \right) \left( \sum_{i=1}^{n} Y_i \right)}{\sqrt{n \left( \sum_{i=1}^{n} X_i^2 \right) - \left( \sum_{i=1}^{n} X_i \right)^2} \sqrt{n \left( \sum_{i=1}^{n} Y_i^2 \right) - \left( \sum_{i=1}^{n} Y_i \right)^2}}
\]
Using this formula was a lot of work. Today you can easily calculate the coefficient of correlation with your calculator. Start appreciating your calculator!!! But you want to make sure that you understand what this coefficient means. Values of r range from -1 to 1. A value of zero shows absolutely no linear correlation between the values of x and y. The sign of r will be the same as the sign of the slope of the best fit line. A value of a positive one or a negative one shows a perfect correlation between x and y, meaning that the data pairs fall exactly on a line. A value of r from .5 to 1 and -.5 to -1 shows a strong relationship.

Be careful that you don’t confuse the ideas of correlation and causation. There may exist a strong correlation between two sets of data, but this does not necessarily imply causation.

**IN OTHER WORDS, CORRELATION DOES NOT IMPLY CAUSATION!!!**

In the case of the Challenger tragedy, there were many “lurking” variables. Some of the “lurking” variables include the following:

- the use of a new type of putty to seal the O-Rings since the putty formerly used was removed from the market by NASA because it contained asbestos
- launch delays
- internal ice changing the frequency of vibration