Math 146 **10.1 - Correlation of Bivariate Data**

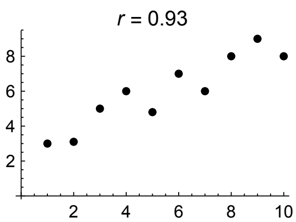
Correlation is a statistical technique that can help determine if there is a relationship or dependency between to sets of paired data, e.g., is there a significant correlation between a persons height and foot length?

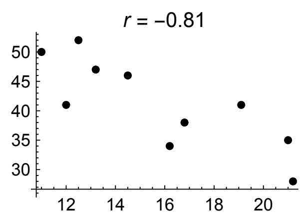
DEFINITIONS

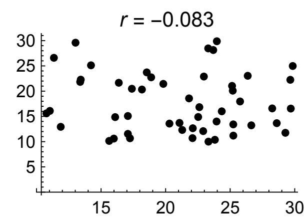
A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

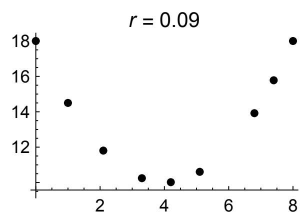
A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

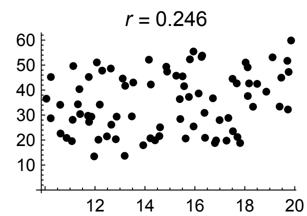
Correlation is measured using the **coefficient of correlation**, ***r***, and can have values between −1 and 1 inclusive.











The correlation coefficient *r* measures the strength of correlation. If the bivariate data appears to be linear, and there are no significant outliers, the correlation coefficient can be calculated using the formula:

$$r = \frac{n(\sum x\,y)-(\sum x)\,(\sum y)}{\sqrt{n(\sum x^2)-(\sum x)^2}\cdot\sqrt{n(\sum y^2)-(\sum y)^2}}$$ (Our calculator will do this for us.)

EXAMPLE 1 Enter the data below into L_1 and L_2 and calculate **2-Var Stats** on the TI84. Write down: n, $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$. Use the above formula to calculate the correlation coefficient. Also, verify that the data show a linear relation. Is there strong correlation between *x* and *y*?

x	1	2	3	4	5
y	2	4	5	8	8

To determine if there is significant correlation between the data we can use a hypothesis test for the population correlation ρ (rho) with $H_0 : \rho = 0$, and $H_1 : \rho \neq 0$, i.e., the population correlation is equal to 0 (no correlation) or the population correlation is not 0, that is, there is correlation. Table **A-6** gives the critical values for correlation for $\alpha = 0.05$ and $\alpha = 0.01$, and is dependent on the number of data points.

EXAMPLE 2 Determine if there is significant correlation for *Example 1*, and for the fifth scatterplot above where $n = 80$, using Table A-6.

P-Value Method for Hypothesis Test for Linear Correlation

For the hypotheses: $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$, the test statistic is: $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ with $n - 2$ degrees of freedom. The P-value can be computed with the TI calculator using **LinRegTTest**, or using ***Correlation and Regression*** with EasyCalc.

EXAMPLE 3 Find the P-value for the data from Example (1), and determine if there is significant correlation.

Interpretation of Correlation

When we conclude there is correlation, what does that mean? The value of $\boldsymbol{r^2}$ is the proportion (or percentage) of variation in y that is explained by the variation in x and the relation between x and y. Suppose the correlation between foot length and height is $r = 0.81$. This means, $0.81^2 \approx 0.66$, or 66% of the variation in height can be explained by the variation in foot length. The other 34% of variation in the height is due to other factors.

Common Errors Involving Correlation

- Correlation DOES NOT imply causation.

(1) The median cost of homes -vs- the number of pirates for the last 200 years is $r \approx -0.86$.
(2) The number of drunk driving accidents and the number of K-12 school teachers over the last 7 years is $r \approx 0.73$.

- Lurking variables can affect the variation in both variables in a study, but are not included in the study.
- If there is no linear correlation, there still may be a strong correlation, just not linear.