10.4 Measuring Correlation

Up to this point we have used a scatter diagram to see if there is a relationship (correlation) between two variables.

We have determined it as positive or negative; zero if there is no correlations. We have called the correlation weak, moderate or strong. We have found the equation of the regression line and used the line for prediction purposes.

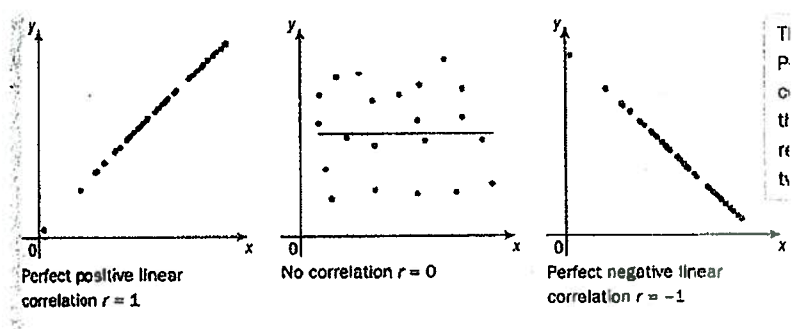
Now we will seek to classify the strength of the correlation numerically.

- There are several scales that are in use;
- We will study a correlation coefficient developed by Karl Pearson.
  - Pearson found the world's first university statistics department at University College London in 1911.

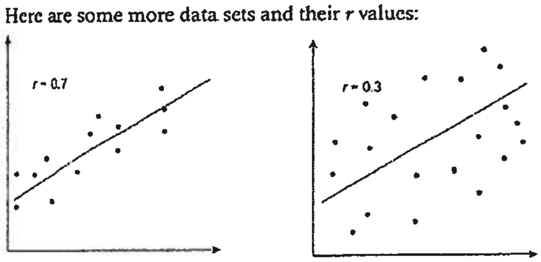
> The **Pearson product-moment correlation coefficient** (denoted by r) is a measure of the correlation between two variables X and Y, giving a value between $+1$ and -1 inclusive. It is widely used in sciences as a measure of the strength of **linear** dependence between two variables.

highest : 1 lowest : -1

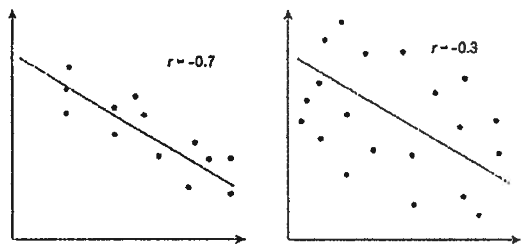
- The regression coefficient is used to determine how nearly the points fall on a straight line or how nearly linear they area
- A perfect correlation will have a regression coefficient of $r = 1.000$.
- Normally in the physical sciences we would like to have a "confidence level" of 0.01 or better.
  - That mean that a coefficient of $r = 0.990$ or higher.



Here are some more data sets and their r values:



For negative correlation, the value of r is also negative.



In the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The r-value from Pearson's correlation coefficient indicates the strength of the relationship between two data sets.

The formula for finding Pearson's correlation coefficient is

$$r = \frac{S_{xy}}{S_x S_y}, \text{ where}$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} \text{ and}$$

$$S_x = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}, \; S_{y=}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}$$

A quick way to interpret the r-value is:

r-value	Correlation
$0 < \|r\| \le 0.25$	Very weak
$0.25 < \|r\| \le 0.5$	Weak
$0.5 < \|r\| \le 0.75$	Moderate
$0.75 < \|r\| \le 1$	Strong

Example: Sue wants to determine the strength of the correlation between the number of spoons of plant food she uses and the extra number of orchids grown from a plant. Use Pearson's correlation coefficient formula to interpret the relationship.

Plant	Spoons of plant food x	Increase in the number of orchids y
A	1	2
B	2	3
C	3	8
D	4	7

$S_{xy} = 60 - \frac{10(26)}{4} = 10$

$S_x = \sqrt{30 - \frac{100}{4}} = \sqrt{5}$

$S_y = \sqrt{126 - \frac{400}{4}} = \sqrt{26}$

$r = \frac{S_{xy}}{S_x S_y} = \frac{10}{\sqrt{5}\sqrt{26}} = .877$ strong positive

$\boxed{\text{Exercise 10F}}$