Section 9.1, Correlation

A **correlation** is a relationship between two quantitative variables. The data can be represented by the ordered pairs (x, y), where x is the **independent (explanatory) variable** and y is the **dependent (response) variable**. **Positive linear correlation** occurs if y tends to increase as x increases. **Negative linear correlation** occurs if y tends to decrease as x increases. If there is no linear trend, we say that there is no linear correlation (Note: There are other kinds of nonlinear correlation, which can happen when the data follow a general trend that is not linear.).

The **correlation coefficient** r is a way to calculate the strength of a linear relationship, as well as whether the correlation is negative or positive:

$$r = \frac{n\sum(xy) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum(x^2) - (\sum x)^2}\sqrt{n\sum(y^2) - (\sum y)^2}}$$

A few facts about r:

- $-1 \leq r \leq 1$.
- If $r > 0$, there is positive linear correlation.
- If $r < 0$, there is negative linear correlation.
- If $r = 1$ or -1, there is a perfect linear relationship between x and y (all data points falls on a line).
- If $r = 0$, there is no linear correlation.
- The further r is from zero, the stronger the linear relationship.

Examples
State whether each of the following data sets has positive or negative linear correlation (or neither). Also calculate the correlation coefficient for each of the following:

1. The number of officers on duty in a Boston city park and the number of muggings for that day are:

Officers	Muggings
10	5
15	2
16	1
1	9
4	7
6	8
18	1
12	5
14	3
7	6

We should choose x to be the number of officers and y to be the number of muggings. After sketching a scatter plot, we see that there is a negative linear correlation.

Calculating, we get that $\sum x = 103$, $\sum y = 47$, $\sum xy = 343$, $\sum x^2 = 1347$, and $\sum y^2 = 295$. Using our formula for r, we get that

$$r = \frac{10 \cdot 343 - 103 \cdot 47}{\sqrt{10 \cdot 1347 - 103^2}\sqrt{10 \cdot 295 - 47^2}} = -0.969$$

2. The age of a Shetland pony (in months) and the average weight of a pony (in kilograms) is:

Age	Weight
3	60
6	95
12	140
18	170
24	185

Let x represent the age of the pony, and y represent its weight. There is a positive linear correlation between x and y. We can calculate that $\sum x = 63$, $\sum y = 650$, $\sum xy = 9930$, $\sum x^2 = 1089$, and $\sum y^2 = 95350$. Then,

$$r = \frac{5 \cdot 9930 - 63 \cdot 650}{\sqrt{5 \cdot 1089 - 63^2}\sqrt{5 \cdot 95350 - 650^2}} = 0.972$$

3. The global average temperature (in degrees Celsius), and number of pirates are:

Temperature	Pirates
14.2	35000
14.4	45000
14.5	20000
14.8	15000
15.1	5000
15.5	400
15.8	17

The relationship has negative linear correlation, since the number of pirates decreases as the temperature increases. We will let x represent the temperature, and y represent the number of pirates. Then, $\sum x = 104.3$, $\sum y = 120417$, $\sum xy = 1738968.6$, $\sum x^2 = 1556.19$, and $\sum y^2 = 3900160289$. Therefore,

$$r = \frac{7 \cdot 1738968.6 - 104.3 \cdot 120417}{\sqrt{7 \cdot 1556.19 - 104.3^2}\sqrt{7 \cdot 3900160289 - 120417^2}} = -0.887$$

Correlation and Causation

Knowing that two variables are correlated does not necessarily imply a cause-and-effect relationship between the variables. Normally, more research is necessary to determine if there is a causal relationship. While it is possible for one variable to affect another, it is also possible that there is a third (confounding) variable on which they both depend. For example, there is high correlation between the number of ice cream sales and the number of drownings on a given day. This does not mean that eating ice cream causes drownings. Since both are common in the summer, especially on hot days, both increase as temperatures increase.

It is also possible that the relationship is coincidental (For example, pirates and temperatures).