The correlation gives a measure of the linear association between two variables. It is a coefficient that does not depend on the units that are used to measure the data and is bounded between -1 and 1. In this class we will consider the following points:

- Scatterplots
- Defining and computing the correlation
- The effect of changing the variables' units
- The effect of non-linear trends and outliers
Correlation

How do you study the relationship between two variables? Pearson considered the data corresponding to the heights of 1,078 fathers and their son's at maturity. The plot appears on page 120 of the textbook (next page in this presentation). A list of these data is difficult to understand, but the relationship between the two variables can be visualized using a scatter diagram, where each pair father-son is represented as a point in a plane. The x-coordinate corresponds to the father's height and the y-coordinate to the son's. We observe that the taller the father the taller the son, as a general tendency. This corresponds to a positive association. In this example we consider the height of the father as an independent variable and he height of the son as a dependent variable.
Figure 1. Scatter diagram for the heights of 1,078 fathers and sons, showing the positive association between son’s height and father’s height. Families where the height of the son equals the height of the father are plotted along the 45-degree line $y = x$. Families where the father is 72 inches tall (to the nearest inch) are plotted in the vertical strip.
The figure shows the scatter diagram of the temperature in 58 locations in Northern California on August 1950. These data are plotted against the elevations of the station where the measurement was recorded. We can see that there is a tendency for locations at high elevations to have lower temperatures compared to those at low elevations.
The correlation coefficient

We have seen that the mean and the standard deviation provide a description of the behavior of a sample for a given variable. When we consider two variables we can calculate the mean and the standard deviation of each of them, but none of those four numbers will give a measure of the association between the two variables.

The correlation coefficient gives a measure of the linear association of two variables

That is, when the scatter plot of the two variables is very close to the straight line we have a correlation that is close to one. A zero correlation corresponds to a diagram where the data are widely scattered around the line.

The correlation coefficient is usually denoted by r and takes values between -1 and 1
The correlation coefficient

A negative coefficient means that the data are clustered around lines with a negative slope. That is, as one variable increases, the other one decreases.

The closer \( r \) is to 1 the stronger the positive linear association between the variables.

The closer \( r \) is to -1 the stronger the negative linear association between the variables.

When \( r \) is equal to 1 or -1 there is total linear association between the variables, this implies that all points lie on a line.

The slope of the line is positive for \( r = 1 \) and negative for \( r = -1 \).
Computing the correlation coefficient

The procedure to compute the correlation coefficients is the following
1. Convert each variable to standard units
2. Calculate the average of the products
The result is the correlation coefficient. The formula is given by
\[ r = \text{average of } (x \text{ in standard units } \times y \text{ in standard units}) \]

**Example:** Consider the data

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>1</td>
<td>13</td>
</tr>
</tbody>
</table>

Each pair of numbers in the table corresponds to a subject.
1. Convert \( x \) to standard units. The average of the \( x \)-values is 4, the SD is 2.
Computing the correlation coefficient

2. Convert \( y \) to standard units. The mean of the \( y \)-values is 7 and the SD is 4.

3. Compute the products of the standard units of the \( x \)-values and the \( y \)-values.

\[
\begin{array}{ccccc}
0.75 & -0.25 & 0.00 & -0.75 & 2.25 \\
\end{array}
\]

4. Take the average of the products

\[
r = \frac{0.75 - 0.25 + 0.00 - 0.75 + 2.25}{5} = 0.40
\]
Example

Consider the data in the table

<table>
<thead>
<tr>
<th>Length</th>
<th>10.7</th>
<th>11.0</th>
<th>9.5</th>
<th>11.1</th>
<th>10.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width</td>
<td>5.8</td>
<td>6.0</td>
<td>5.0</td>
<td>6.0</td>
<td>5.3</td>
</tr>
<tr>
<td>Length</td>
<td>10.7</td>
<td>9.9</td>
<td>10.6</td>
<td>10.0</td>
<td>12.0</td>
</tr>
<tr>
<td>Width</td>
<td>5.8</td>
<td>5.2</td>
<td>5.7</td>
<td>5.3</td>
<td>6.3</td>
</tr>
</tbody>
</table>

These correspond to the lengths and the widths of 10 chitons. We have that the averages are 10.58 for the lengths and 5.64 for the widths. The standard deviations are 0.6734 and 0.3980 respectively. Thus the transformation to standard units yields
Then \( r = 0.969 \) is the average of the products. This implies that there is a strong positive linear association between the length and the width of the chitons.
Positive Correlation

Suppose that this scatter diagram corresponds to the standard units of the data. The signs of the quadrants are determined by the sign of the products. There are more points in the positive quadrants than in the negative quadrants. So the average is positive yielding a positive correlation.
Negative Correlation

In this case there are more points in the negative quadrants than in the positive quadrants. So the average is negative, yielding a negative correlation.
Features of the correlation coefficient

Suppose you measure the correlation between the temperature in New York and that in Boston during a month. Do you expect to get the same correlation if you measure it in Celsius than if you measure it in Fahrenheit? Recall the process of obtaining the correlation. The first step is to convert the samples of both variables to standard units. This implies that the correlation does not depend on units. Thus, no matter if you use Celsius or Fahrenheit, you will get the same correlation.

On the other hand, since the correlation depends on the product of the two variables, it does not matter whether you consider the correlation the temperature in NY and that in Boston or the correlation between the temperature in Boston and that in NY.
Features of the correlation coefficient

The correlation coefficient has the following properties:

• The correlation is not affected when the two variables are interchanged.
• The correlation is not changed if the same number is added to all the values of one of the variables.
• The correlation is not changed if all the values of one of the variables is multiplied by the same positive number. It will change sign if the number is negative.
Non linear association

Correlation is useful only when measuring the degree of linear association between two variables. That is, how much the values from two variables cluster around a straight line. The variables in this plot have an obvious non-linear association.

Nevertheless the correlation between them is 0.3. This is because the points are clustered around a curve that has the shape of a parabola and not a straight line.
Similarly, the presence of outliers produces an artificially low correlation between variables that have a high degree of linear association. The correlation of the data in the plot is 0.944. When the points marked with crosses in red are considered, the correlation drops to 0.75.
Examples

Find the correlation coefficient for each of the following data sets:

Example 1:

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>y</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The average of $x$ is 2, the SD is 1. The average of $y$ is 3 and the SD is 2. Then, the standardized values are

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>-1</th>
<th>-1</th>
<th>-1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$x \times y$</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
</tr>
</tbody>
</table>

and the average of the last column is -0.80.
Example 2:

\[ x \]
\[
\begin{array}{cccccccc}
1 & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 4 \\
1 & 2 & 1 & 3 & 1 & 4 & 1 & 2 & 2 & 3 \\
\end{array}
\]

\( x \) has not changes, \( y \) has average equal to 2 and a SD equal to 1. Then, the standardized values are

\[
\begin{array}{cccccccc}
x & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 2 \\
y & -1 & 0 & -1 & 1 & -1 & 2 & -1 & 0 & 0 & 1 \\
x \times y & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 2 \\
\end{array}
\]

and the average of the last column is 0.30.
Example 3:

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

We observe that $y = 2 \times x$ and so there is total linear association between $x$ and $y$, implying that the correlation is 1.
Ecological correlations

Correlations based on rates or averages can be misleading.

**Example 1:** Relationship between the rate of cigarette smoking (per capita) and the rate of deaths from lung cancer in 11 countries gave correlation 0.7. However, it is not countries which smoke and get cancer, but people. To measure the strength of the relationship for people, we must have data for individual people.
Ecological correlations

Example 2: From Current Population Survey data for 1993, you can compute the correlation between income and education for men age 25-54 in US: $r=0.44$. For each state you can compute the average educational level and income. Finally, you can compute the correlation between the pairs of averages: this is $0.64$!! If you use the correlation for the states to estimate the correlation for the individuals, you would be way off. The reason is that within each state, there is a lot of spread around the averages. Replacing the states by their averages eliminates the spread, and gives a misleading impression of tight clustering.
Ecological correlations

Ecological correlations are based on rates or averages. They are often used in political science and sociology. And they tend to overstate the strength of an association.
For school children, shoe size is strongly correlated with reading skills. However, learning new words does not make the feet get bigger. Instead, there is a third factor involved-age. As children get older, they learn to read better and they outgrow their shoes.

Correlation measures association. But association does not necessarily show causation. It may only show that both variables are simultaneously influenced by some third variable.