Section 10.2
Correlation between two variables (x and y)

Objective
Investigate how two variables (x and y) are related (i.e. correlated). That is, how much they depend on each other.

Definitions
A correlation exists between two variables when the values of one appears to somehow affect the values of the other in some way.

In this class, we are only interested in linear correlation

Linear correlation coefficient : r
A numerical measure of the strength of the linear relationship between two variables, x and y, representing quantitative data.

r always belongs in the interval (-1,1) (i.e. \(-1 \leq r \leq 1\))

We use this value to conclude if there is (or is not) a linear correlation between the two variables.

Exploring the Data
We can often see a relationship between two variables by constructing a scatterplot.

Positive Correlation
We say the data has positive correlation if the data follows a line (with a positive slope).

The correlation coefficient (r) will be close to +1

Negative Correlation
We say the data has negative correlation if the data follows a line (with a negative slope).

The correlation coefficient (r) will be close to -1
We say the data has **no correlation** if the data does not seem to follow any line. The correlation coefficient \( r \) will be close to 0.

**Interpreting \( r \)**

- \( r \approx 1 \) **Strong positive** linear correlation
- \( r \approx 0 \) **Weak** linear correlation
- \( r \approx -1 \) **Strong negative** linear correlation

**Nonlinear Correlation**

The data may follow a curve, but if the data is not **linear**, the **linear correlation coefficient** \( r \) will be close to zero.

**Requirements**

1. The sample of paired \((x, y)\) data is a **random sample** of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a **straight-line pattern**.
3. The **outliers must be removed** if they are known to be errors. *(Note: We will not do this in this course)*

**Notation**

- \( n \) Number of pairs of sample data
- \( \Sigma \) Denotes the addition of the items
- \( \Sigma x \) The sum of all \( x \)-values \( \Sigma x = x_1 + x_2 + \ldots + x_n \)
- \( \Sigma y \) The sum of all \( y \)-values \( \Sigma y = y_1 + y_2 + \ldots + y_n \)
- \( \Sigma x^2 \) The sum of the squares for all \( x \)-values \( \Sigma x^2 = x_1^2 + x_2^2 + \ldots + x_n^2 \)
- \( (\Sigma x)^2 \) The sum of the \( x \)-values, then squared \( (\Sigma x)^2 = (x_1 + x_2 + \ldots + x_n)^2 \)
- \( \Sigma xy \) The sum of the products of \( x \) and \( y \) \( \Sigma xy = x_1y_1 + x_1y_2 + \ldots + x_ny_n \)
Correlation Coefficient

- **Sample** linear correlation coefficient
- **Population** linear correlation coefficient

\( r \) measures the strength of a linear relationship between the paired values in a sample.

\[
r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2} \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}
\]

We use **StatCrunch** to compute \( r \) (Don't panic!)

Example 1a

Make a scatterplot for the heights of mother, daughter

Enter data on **StatCrunch** (Mother in 1st column, daughter in 2nd column)

Select var1 for **X variable** (height of mother)
Select var2 for **Y variable** (height of daughter)
Click **Create Graph!**

Voila! (Does there appear to be correlation?)

Example 1b

Find the linear correlation coefficient of the heights

**Stat – Summary Stats – Correlation**
Select var1 and var2 so they appear in the right box
Click Calculate

Find the linear correlation coefficient of the heights

Example 1b

Determining if Correlation Exists
We determine whether a population is correlated via a two-tailed test on a sample using a significance level (α)

\[ H_0: \rho = 0 \quad \text{(i.e. not correlated)} \]
\[ H_1: \rho \neq 0 \quad \text{(i.e. is correlated)} \]

Again, two methods available:

Critical Regions (Use Table A-6)
P-value (Use StatCrunch)

Note: In most cases we use significance level \( \alpha = 0.05 \)

Using Critical Regions
Use Table A-6 to find the critical values, which depends on the sample size \( n \). Use both positive and negative values (two-tailed)

- If the \( r \) is in the critical region, we conclude that there is a linear correlation.
  (Since \( H_0 \) is rejected)
- If the \( r \) is not in the critical region, we say there is insufficient evidence of correlation.
  (Since we fail to reject \( H_0 \))

Table A-6

<table>
<thead>
<tr>
<th>( n )</th>
<th>( a = 0.05 )</th>
<th>( a = 0.01 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.905</td>
<td>0.940</td>
</tr>
<tr>
<td>5</td>
<td>0.818</td>
<td>0.879</td>
</tr>
<tr>
<td>6</td>
<td>0.754</td>
<td>0.816</td>
</tr>
<tr>
<td>7</td>
<td>0.707</td>
<td>0.774</td>
</tr>
<tr>
<td>8</td>
<td>0.666</td>
<td>0.736</td>
</tr>
<tr>
<td>9</td>
<td>0.632</td>
<td>0.708</td>
</tr>
<tr>
<td>10</td>
<td>0.602</td>
<td>0.684</td>
</tr>
<tr>
<td>11</td>
<td>0.576</td>
<td>0.661</td>
</tr>
<tr>
<td>12</td>
<td>0.550</td>
<td>0.641</td>
</tr>
<tr>
<td>13</td>
<td>0.522</td>
<td>0.623</td>
</tr>
<tr>
<td>14</td>
<td>0.497</td>
<td>0.606</td>
</tr>
</tbody>
</table>

From Example 1b, we found \( r = 0.802 \)
Since \( n = 20 \) and \( \alpha = 0.05 \), using Table A-6, we find the critical values to be: 0.444, -0.444
Since \( r \) is in the critical region (reject \( H_0 \)), we conclude the data is linearly correlated (under 0.05 significance).
Using *P*-value

Use *StatCrunch* to calculate the two-tailed *P*-value from a sample set (see Example 1c)

- If the *P*-value is less than α, we conclude that **there is a linear correlation**. (Since H₀ is rejected)
- If the *P*-value is greater than α, we say there is **insufficient evidence of correlation**. (Since we fail to reject H₀)

Example 1c

Use a 0.05 significance level to determine if the heights are linearly correlated.

Using *P*-value

- On *StatCrunch*: Stat → Summary Stats → Correlation
- Select var1, var2 so they appear in right box
- Click Next
- Check “Display two-sides *P*-value from sig. test”
- Click Calculate
- Result: *P*-value < 0.0001

Since *P*-value is less than α=0.05 (reject H₀), we conclude the data is linearly correlated

Caution!

Know that the methods of this section apply only to a **linear correlation**.

If you conclude that there is no linear correlation, it is possible that there is some other association that is not linear.

Rounding the Linear Correlation Coefficient

- Round *r* to **three decimal places** so that it can be compared to critical values in Table A-6

Properties of the Linear Correlation Coefficient *r*

1. -1 ≤ *r* ≤ 1
2. If all values of either variable are converted to a different scale, the value of *r* does not change.
3. The value of *r* is **not affected by the choice** of *x* and *y*. Interchange all *x*-values and *y*-values and the value of *r* will not change.
4. *r* measures **strength** of a linear relationship.
5. *r* is very sensitive to **outliers**, they can dramatically affect its value.

Example 2

A new medication for high blood pressure was tested on a batch of 18 patients with different ages. The results were as follows:

(a) Plot the points
(b) Find the correlation coefficient
(c) Use a 0.05 significance level to test linear correlation
A survey of 15 people was conducted to see how many friends people had on Facebook vs. their shoe size. The results were as follows:

Shoe size: 8.4 9.0 9.1 7.8 9.8 8.3 9.9 8.8 9.1 9.5 9.6 8.1 8.2 8.3 8.0 9.4

(a) Plot the points
(b) Find the correlation coefficient
(c) Use a 0.05 significance level to test linear correlation
**Example 3b** Find Correlation Coefficient

Go to: **Stats → Summary Stats → Correlation**

Select `var1` and `var2`, hit **Calculate**

\[ r = 0.409 \]

**Example 3c** Test for Correlation (\( \alpha = 0.05 \))

Using P-value

Go to: **Stats → Summary Stats → Correlation**

Select `var1` and `var2`, hit **Next**

Check box, hit **Calculate**

\[ P\text{-value} = 0.1297 \]

Since **P-value** greater than \( \alpha = 0.05 \) (fail to reject \( H_0 \)), we conclude there is **no correlation**

**Example 4**

Using the data in **Example 2** (blood pressure vs. age), we found that the linear correlation coefficient is \( r = 0.964 \)

What proportion of the variation in the patients' blood pressure can be explained by the variation in the patients' age?

\[ r = 0.964, \quad r^2 = 0.929 \]

We conclude that 0.929 (or about 93%) of the variation in blood pressure can be explained by the linear relationship between the age and blood pressure.

This implies about 7% of the variation in blood pressure cannot be explained by the age.

**Common Errors Involving Correlation**

1. **Causation**: It is wrong to conclude that correlation implies causality.

2. **Linearity**: There may be some relationship between \( x \) and \( y \) even when there is no linear correlation.
Caution!!!

Know that correlation does not imply causality.

There may be correlation without causality.