The Sample Linear Correlation Coefficient

$r_{XY}$ (or just $r$ for short) is the sample linear correlation coefficient.

$r_{XY}$ measures the strength and direction of linear association between two quantitative variables $X$ and $Y$.

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{s_X s_Y},$$

where $n$ is number of pairs of observations, $\bar{X}$ is the sample average of the $X$ data, $\bar{Y}$ is the sample average of the $Y$ data, $s_X$ is the sample standard deviation of the $X$ data, and $s_Y$ is the sample standard deviation of the $Y$ data.

For example, consider 11 families randomly selected from the population of families with one brother and one sister, both full grown. Let $X_i$ denote the height (in inches) of the brother in the $i$th family. Let $Y_i$ denote the height (in inches) of the sister in the $i$th family.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$X_i$</th>
<th>$Y_i$</th>
<th>$X_i - \bar{X}$</th>
<th>$Y_i - \bar{Y}$</th>
<th>$(X_i - \bar{X})(Y_i - \bar{Y})$</th>
<th>$(X_i - \bar{X})^2$</th>
<th>$(Y_i - \bar{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>71</td>
<td>69</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>64</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>66</td>
<td>65</td>
<td>-3</td>
<td>1</td>
<td>-3</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>67</td>
<td>63</td>
<td>-2</td>
<td>-1</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>65</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>71</td>
<td>62</td>
<td>2</td>
<td>-2</td>
<td>-4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>65</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>73</td>
<td>64</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>72</td>
<td>66</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>59</td>
<td>-4</td>
<td>-5</td>
<td>20</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>11</td>
<td>66</td>
<td>62</td>
<td>-3</td>
<td>-2</td>
<td>6</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>759</td>
<td>704</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>74</td>
<td>66</td>
</tr>
</tbody>
</table>

$$\bar{X} = \frac{759}{11} = 69 \quad \bar{Y} = \frac{704}{11} = 64 \quad S_X = \sqrt{\frac{74}{11-1}} \quad S_Y = \sqrt{\frac{66}{11-1}}$$

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{s_X s_Y} = \frac{3.9}{\sqrt{(7.4)(6.6)}} \approx 0.558$$

$r_{XY}$ estimates the population linear correlation coefficient $\rho_{XY}$.

$r_{XY}$ is dimensionless and is always between -1 and 1.

$r_{XY} = 1$ if and only if all data points fall perfectly on a line with positive slope.

$r_{XY} = -1$ if and only if all data points fall perfectly on a line with negative slope.

$r_{XY} = 0$ means there is no linear association between $X$ and $Y$. 
Write the letter for each pair of variables on the number line to indicate the value of \( r_{XY} \) that you would expect to see.

<table>
<thead>
<tr>
<th>Pair</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Stalk Diameter of Corn Plant</td>
<td>Weight of Corn Plant</td>
</tr>
<tr>
<td>B</td>
<td>Person’s Age</td>
<td>Person’s Year of Birth</td>
</tr>
<tr>
<td>C</td>
<td>Daily Dow Jones Industrial Average</td>
<td>Daily Rainfall in Seattle</td>
</tr>
<tr>
<td>D</td>
<td># of Ultrasounds During Pregnancy</td>
<td>Birth Weight of Baby</td>
</tr>
<tr>
<td>E</td>
<td>U.S. Monthly Ice Cream Cone Sales</td>
<td>Drowning per Month in U.S.</td>
</tr>
<tr>
<td>F</td>
<td>Age of Wife</td>
<td>Age of Husband</td>
</tr>
</tbody>
</table>

\[-1 \quad \quad \quad \quad \quad 0 \quad \quad \quad \quad \quad 1\]