Return to Old Faithful

The goal of this activity is to understand scatter diagrams of pairs of data and the linear correlation coefficient between two variables.

A park ranger is interested in determining if there is a relation between the interval before an eruption and the duration time of the eruption.

The following data set represent the duration times (in minutes) of a sample of eruptions of Old Faithful in Yellowstone Park and the interval between eruptions that immediately preceded each eruption.

<table>
<thead>
<tr>
<th>interval</th>
<th>70</th>
<th>75</th>
<th>73</th>
<th>69</th>
<th>81</th>
<th>73</th>
<th>80</th>
<th>51</th>
<th>62</th>
<th>79</th>
</tr>
</thead>
<tbody>
<tr>
<td>duration</td>
<td>4.1</td>
<td>3.8</td>
<td>4.7</td>
<td>4.3</td>
<td>4.5</td>
<td>3.5</td>
<td>4.4</td>
<td>2.2</td>
<td>1.9</td>
<td>3.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>interval</th>
<th>61</th>
<th>77</th>
<th>57</th>
<th>81</th>
<th>91</th>
<th>77</th>
<th>90</th>
<th>61</th>
<th>75</th>
<th>67</th>
</tr>
</thead>
<tbody>
<tr>
<td>duration</td>
<td>1.8</td>
<td>4.1</td>
<td>1.9</td>
<td>5.1</td>
<td>5.1</td>
<td>4.1</td>
<td>4.7</td>
<td>2.7</td>
<td>3.8</td>
<td>2.6</td>
</tr>
</tbody>
</table>

1. Enter the data above in your calculator. Enter the interval times in L1 and the duration times in L2. Be sure to enter them in order so that the times correspond to each other.

2. Create a scatter plot of the data. Construct the scatter diagram in the graph below.
3. Briefly describe any visible pattern between the interval time and the duration of eruption.

---

**Linear Correlation Coefficient**

The linear correlation coefficient is a measure of the strength of linear relation between two quantitative variables. The formula for the sample correlation coefficient is

\[ r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \]

where \( \bar{x} \) is the sample mean and \( s_x \) is the sample standard deviation of the explanatory variable.

\( \bar{y} \) is the sample mean and \( s_y \) is the sample standard deviation of the response variable.

\( n \) is the number of individuals in the sample.

Notice that the numerator expression in the definition above is just the product of the z-scores of the \( x \) and \( y \) variables. Thus, an alternative definition of the sample correlation coefficient is

\[ r = \frac{\sum z_x \cdot z_y}{n - 1} \]

Recall that the z-score of a value is the position (relative to the standard deviation) of the value with respect to the mean.
Return to Old Faithful

4. Calculate the statistics of the interval and duration. The TI command is the following:

\[ 2 \text{-Var Stats L1,L2} \]

\[ \sum x = \quad \sum y = \quad \min X = \]
\[ \sum x^2 = \quad \sum y^2 = \quad \max X = \]
\[ Sx = \quad Sy = \quad \min Y \]
\[ \sigma x = \quad \sigma y = \quad \max Y = \]
\[ n = \quad \Sigma xy = \]

5. On the graph in item 2, graph a vertical line at \( \text{ } \) and a horizontal line at \( \text{ } \). These lines form four quadrants. Describe the sign of the \( z \)-score (positive or negative) of values in each of the quadrants and the sign of the product of the two \( z \)-scores

<table>
<thead>
<tr>
<th>Quadrant</th>
<th>( z )-score of ( x ) values:</th>
<th>( z )-score of ( y ) values:</th>
<th>Product of the ( z )-scores:</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>( \text{ } )</td>
<td>( \text{ } )</td>
<td>( \text{ } )</td>
</tr>
<tr>
<td>II</td>
<td>( \text{ } )</td>
<td>( \text{ } )</td>
<td>( \text{ } )</td>
</tr>
<tr>
<td>III</td>
<td>( \text{ } )</td>
<td>( \text{ } )</td>
<td>( \text{ } )</td>
</tr>
<tr>
<td>IV</td>
<td>( \text{ } )</td>
<td>( \text{ } )</td>
<td>( \text{ } )</td>
</tr>
</tbody>
</table>

6. Determine the number of points in each of the quadrants. How many products of the \( z \)-scores should be positive and how many should be negative? Based on these values, do you think the sum of the products should be positive, negative, or close to zero?
7. Enter the \( z \)-scores of the interval times in L3. To do this, you may highlight L3 and enter the command \((L1-|)/Sx\). Next, enter the \( z \)-scores of the duration times in L4.

8. Calculate the product of the \( z \)-scores in L5. How many values in L5 are positive and how many are negative? Does this agree with your answer to item 6?

9. Use the \texttt{sum(L5)} command to add the products of the \( z \)-scores. Divide the result by \( n - 1 \). Write this value of \( r \) below.

\[
r = \frac{\sum z_x \cdot z_y}{n - 1}
\]

10. An equivalent formula for the sample correlation coefficient is given below. Use this formula to verify the calculation of \( r \).

\[
r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \cdot \sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}}
\]

Please store the data lists: L1\rightarrow OFINT and L2\rightarrow OFDUR.