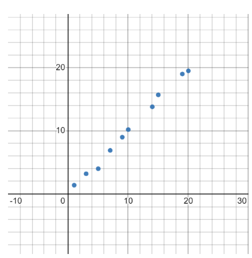
9.2: Quantifying Direction and Strength

In the last lesson, we explored the direction and strength of relationships between two quantitative variables. Now we will begin to explore how to model these relationships. When two variables are related, we say that they **correlate**, and that there is **correlation** between them. Some relationships between the variables in scatterplots can be summarized well by a line. We call such relationships **linear**. Other relationships are better summarized by a curve rather than a line. We call these **nonlinear** or **curvilinear**. Recall that the direction of a relationship can be either positive or negative. Lines and curves can often be fit to the data in a scatterplot. A small amount of deviation away from some line or curve means that the explanatory variable is a good predictor for values of the response variable and we say the relationship is strong. A large amount of deviation from the line or curve indicates that the relationship is weak.

Linear Correlation

One measure of the linear correlation between two variables is called the **linear correlation coefficient**. This correlation coefficient is represented with the letter r. We will use desmos to calculate the value of r for a given scatterplot. Given below is a dataset accompanied by its corresponding scatterplot.

x1	y1
1	1.4
3	3.2
5	4
7	6.9
9	9
10	10.2
14	13.8
15	15.7
19	19
20	19.5



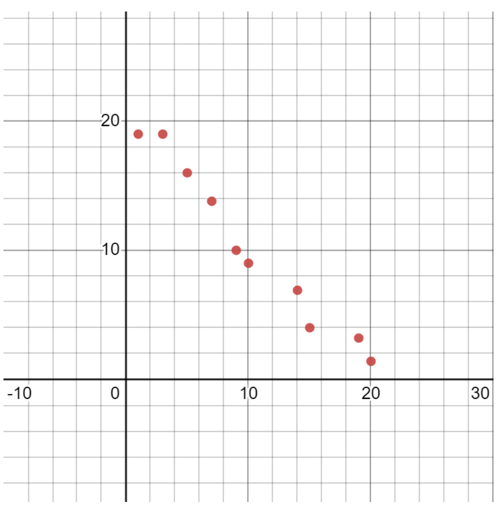
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

To find the linear correlation coefficient,

1. Go to https://www.desmos.com/calculator.
2. Copy and paste the data set above into line 1, or click the plus icon in the top left corner of the calculator, select table, and enter the values into the table.
3. To find r, type $\text{corr}(x1, y1)$ into line 2.
    a. For the data above, $r = \text{corr}(x1, y1) \approx 0.997$

1. Find the linear correlation coefficient for the following two datasets. The corresponding scatterplots are provided. Round r to three decimal places.
   a. $r = \text{corr}(x2, y2) \approx$ _____

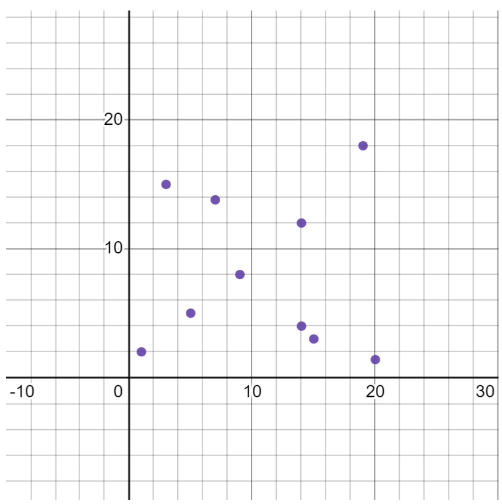
x2	y2
1	19
3	19
5	16
7	13.8
9	10
10	9
14	6.9
15	4
19	3.2
20	1.4



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

b. $r = \text{corr}(x3, y3) \approx$ ______

x3	y3
1	2
3	15
5	5
7	13.8
9	8
14	4
14	12
15	3
19	18
20	1.4

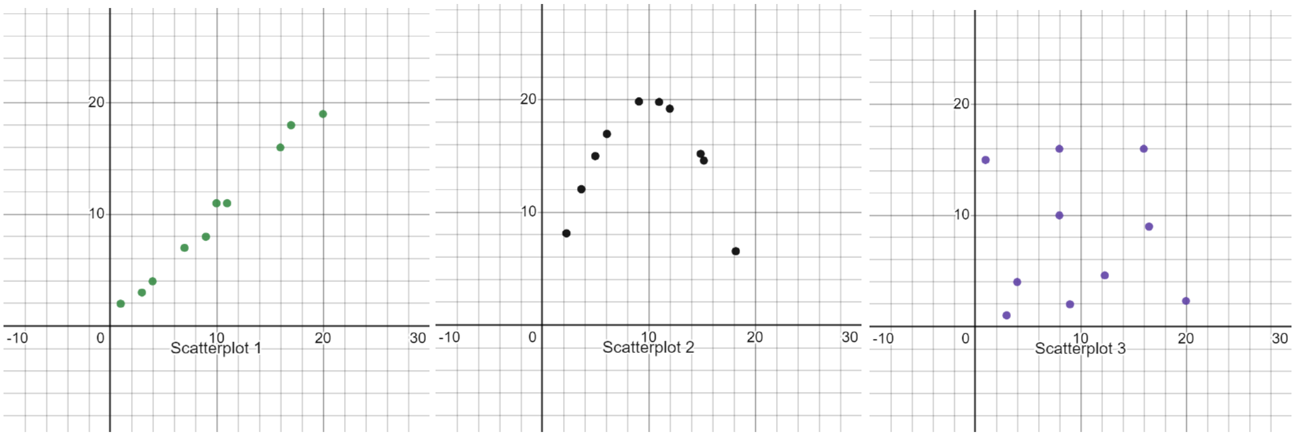


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

2. What do you think the linear correlation coefficient (r) measures?

3. What is the largest possible value of r? What is the smallest possible value for r?

4. Consider the following scatterplots:

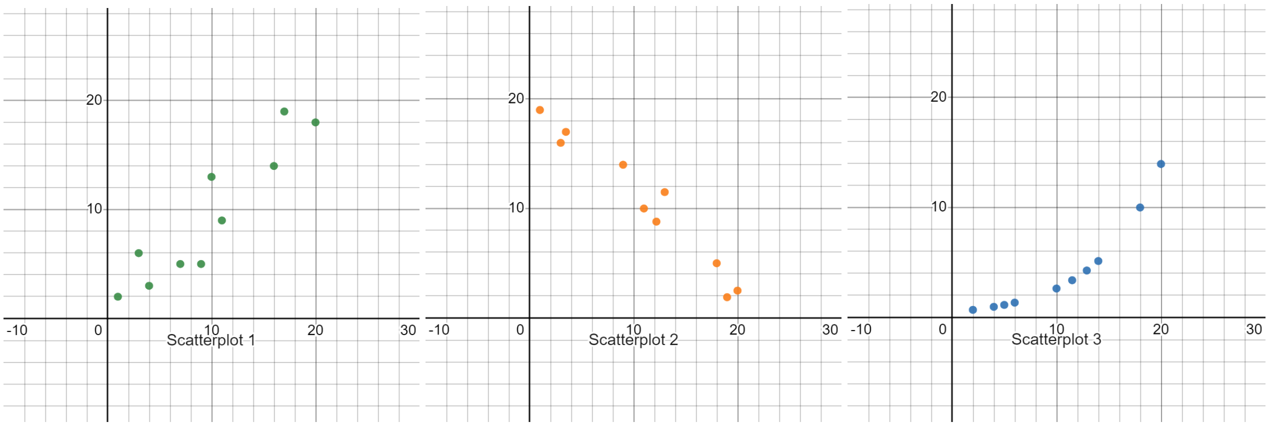


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Decide which two of these scatterplots have an r-value close to 0 without calculating the r value. Explain why you think this.

b. Do both scatterplots with an r-value close to 0 have a weak relationship? If not, explain why.

5. Consider the following scatterplots:



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Determine which two scatterplots above have an r-value close to 0.92 without calculating the r-value. Explain why you think this.

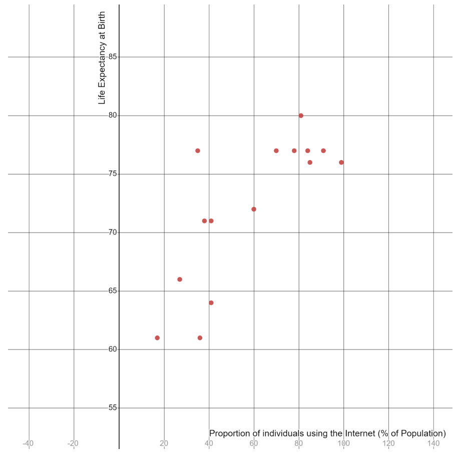
b. Can variables have a nonlinear relationship when r is close to 1? Explain why you think this.

Characteristics of the Linear Correlation Coefficient

- r is between -1 and 1 (inclusive). A scatterplot with an r-value close to -1 or 1 has a strong linear association. Scatterplots with r-values close to 0 have weak linear associations.
- The linear correlation coefficient, r, is a number that describes the direction and strength of a scatterplot with a linear association.
- The sign of r (positive or negative) indicates the direction of a scatterplot with a linear association.

The r-value only gives us reliable information about direction and strength of *linear associations* and should not be used to quantify patterns of a nonlinear association.

6. The scatterplot below shows the rate of individuals using the internet (% of population) and the life expectancy at birth for 15 countries in 2020[21].



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. In this group of 15 countries, does an increase in internet usage tend to be associated with an increase or decrease in life expectancy?

b. The pattern in the scatterplot above indicates a fairly strong, positive, *nonlinear* relationship. Based on this observation, someone might suggest that one way to improve a country's life expectancy would be to get more people online. Is this a reasonable conclusion? Explain.

c. Is the relationship between internet usage and life expectancy one of cause and effect? Consider the type of study that was performed to collect the data.

In an observational study where there is a relationship between two variables, we cannot conclude that the relationship is one of cause and effect. Such conclusions are determined in experimental studies. Sometimes, there is a third variable, possibly unconsidered, that drives changes for both the explanatory and response variables. This additional variable is called a **lurking variable**.

d. Suggest a possible lurking variable that might explain the relationship in internet usage and life expectancy in the data above.

Reference

[21] *World Bank Group*. World Bank. (n.d.). Accessed July 13, 2022, from https://www.worldbank.org/en/home

This page titled 9.2: Quantifying Direction and Strength is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.