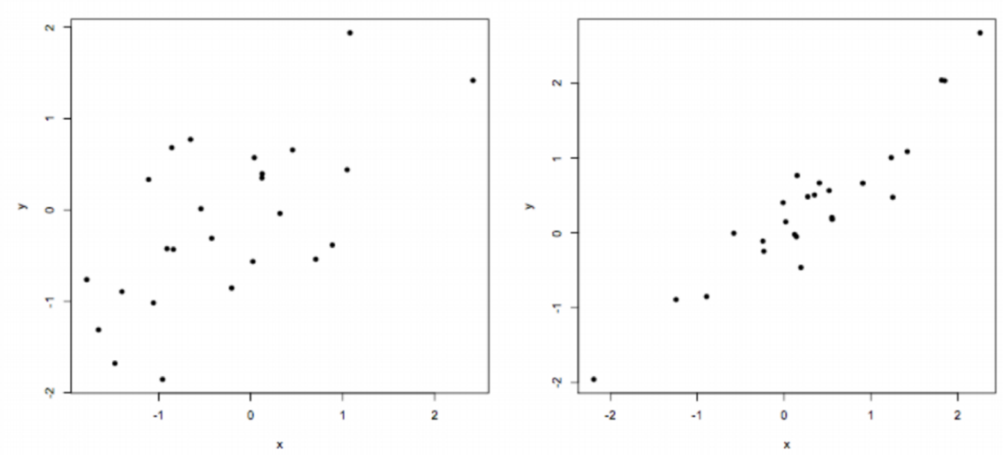
# Chapter Six: Introduction to Correlation

### Review

Describe the form, strength, direction, and whether there are any outliers for the following plots.



You may have noticed last class that the distinction between weak, moderate, and strong are somewhat unclear. A more precise measure of the strength of an association is necessary.

## Correlation

The ________________________________________ measures the strength of a linear association between two quantitative variables. We refer to the correlation coefficient using the mathematical notation $\rho$ or $r$.

### Properties

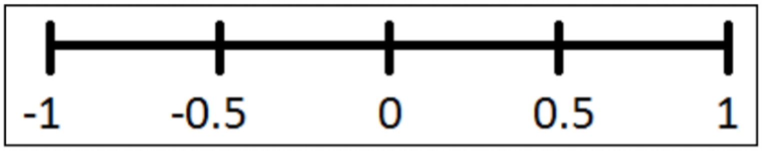
-
-
-
-

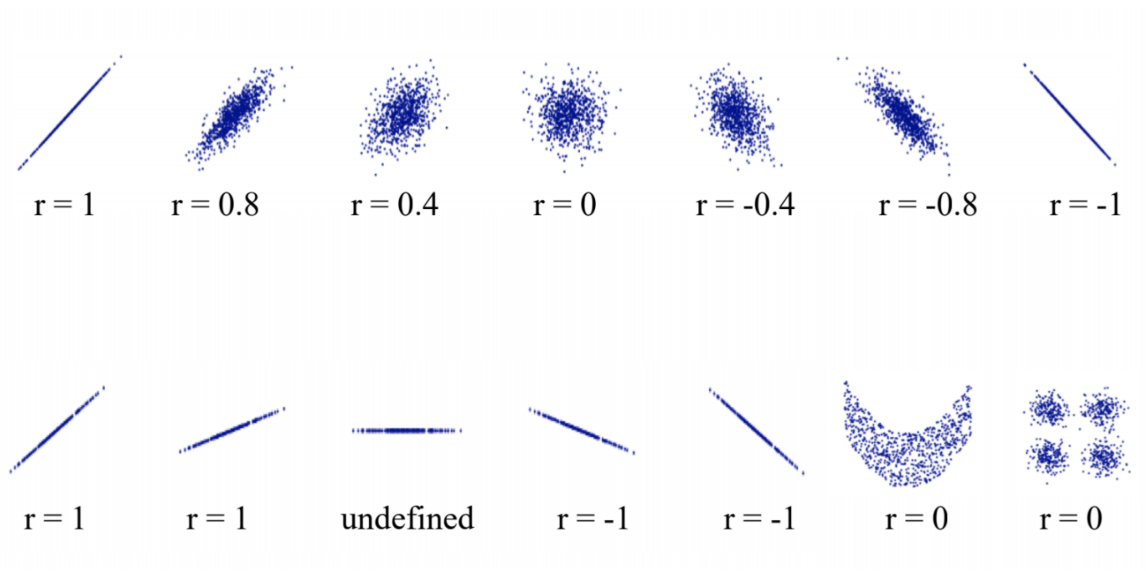
### Association

-
-

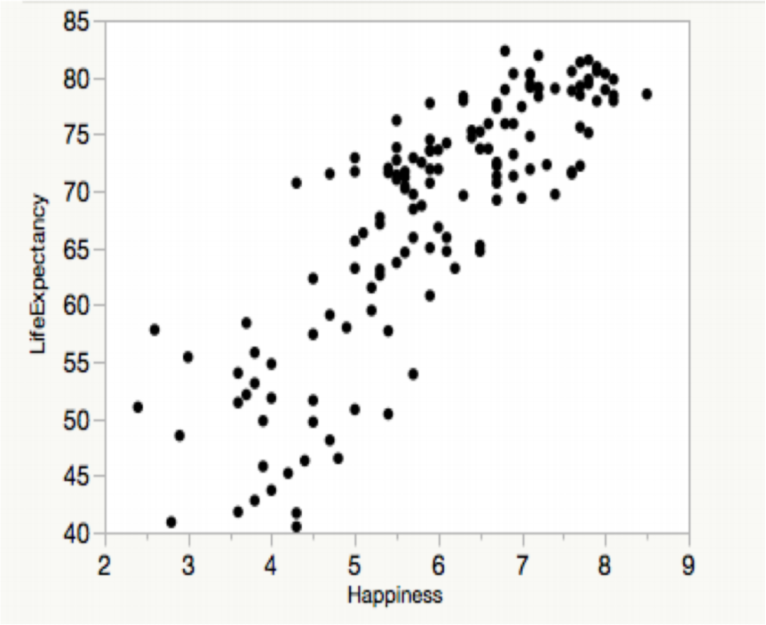
### Strength of Linear Association

-
-
-





As an example, let's return to the Happy planet example we covered on last time. Guess the correlation of the scatterplot.



**Remember:** Since the correlation is unitless, changing the scale of variables to either x or y will not change the correlation.

Please see go to this link: http://www.rossmanchance.com/applets/GuessCorrelation.html

and try guessing the correlation for a few examples of data.

For two quantitative variables $x$ and $y$, we calculate the correlation as follows:

$$\rho = \frac{1}{n-1}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}\right)$$

where:

- $n$ :
- $\bar{x}$ :
- $\bar{y}$ :
- $s_x$ :
- $s_y$ :

Typically, if you are asked to compute the correlation you will be given $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $s_x, s_y$. Still helpful to be able to know how to calculate correlation by hand!

**Example** *Let's reconsider the gas-milead versus horse-power example that we studied last time. Suppose we have:*

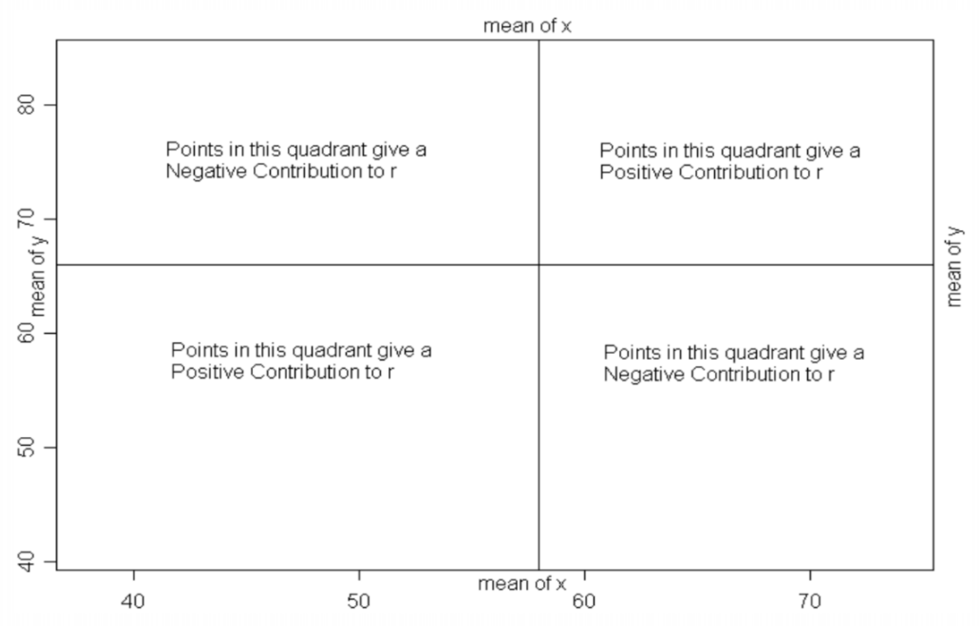
$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -5154.31$$
$$s_x = 66.28 \quad s_y = 6.39$$

*What is the correlation?*

**Example** *Suppose we have the following two quantitative variables, calculate the correlation between these two variables.*

| x | y |
|---|---|
| 1 | 4 |
| 2 | 8 |
| 3 | 7 |
| 4 | 8 |
| 5 | 10 |
| 6 | 12 |

*As you can guess from the above equation, the mean of both the $x$ and $y$ variable influence the correlation. If we were to plot the averages $x$ and $y$ we could visualize how the relationship works.*

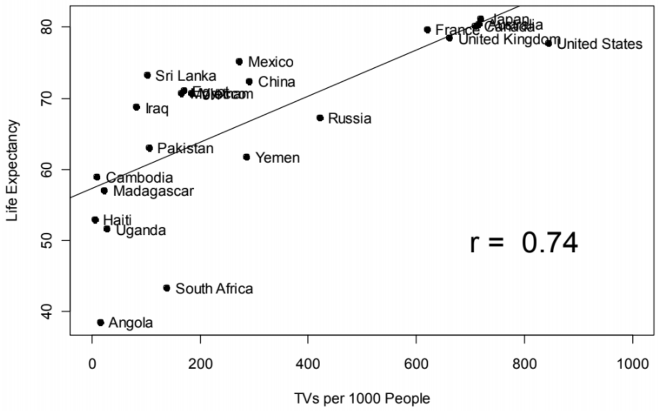

mean of x
Points in this quadrant give a Negative Contribution to r
Points in this quadrant give a Positive Contribution to r
Points in this quadrant give a Positive Contribution to r
Points in this quadrant give a Negative Contribution to r
mean of y
mean of y
mean of x
40
50
60
70
80
40
50
60
70

**Note:** Remember! Association is not Causation! Association between an explanatory and response variable ________________________ mean that the changes in explanatory variables ________________________ changes in the response variable.

Examples of spurious correlation: https://www.tylervigen.com/spurious-correlations

## Lurking Variables

**Example** *Many news stories focus on association not causation, take for example a study that gained some media attention a number of years ago about owning a TV and life expectancy.*



The nature of data necessitates that occasionally there are variables that effect results of our studies that we may not be accounting for. Variables that can impact the relationship between two variables of interest are called ______________________________________________. Typically a lurking variable is related to ______________________ the explanatory and the response variable. For each of the following examples, state what lurking variable(s) could influence these situations.

- High correlation between number of cases of respiratory infections and days of low wind speeds.

- High correlation between number of shark attacks and icecream sales.

- High correlation between road salt purchased and number of car accidents.

So how can we establish causation? We can establish causation by performing ______________________ which we will talk more about in Chapter 11.

**Characteristics of Experiments**

-

-

For each scenario, can we establish causation?

- Testing the relationship between water level and growth of corn. Corn is grown in similar light, soil, and temperature conditions in a green house.
- Opt-in online survey of consumer preferences and sales for a clothing business

- Review of historical records to find an association between grain yields and deaths by starvation
- Studying the relationship between smoking and prenatal health. Participants are selected along similar demographic categories in the same region.