Relationships

Stephen B. Holt, Ph.D.

Rockefeller College
of Public Affairs & Policy
University at Albany State University of New York

February 8, 2022

Two Variable Example

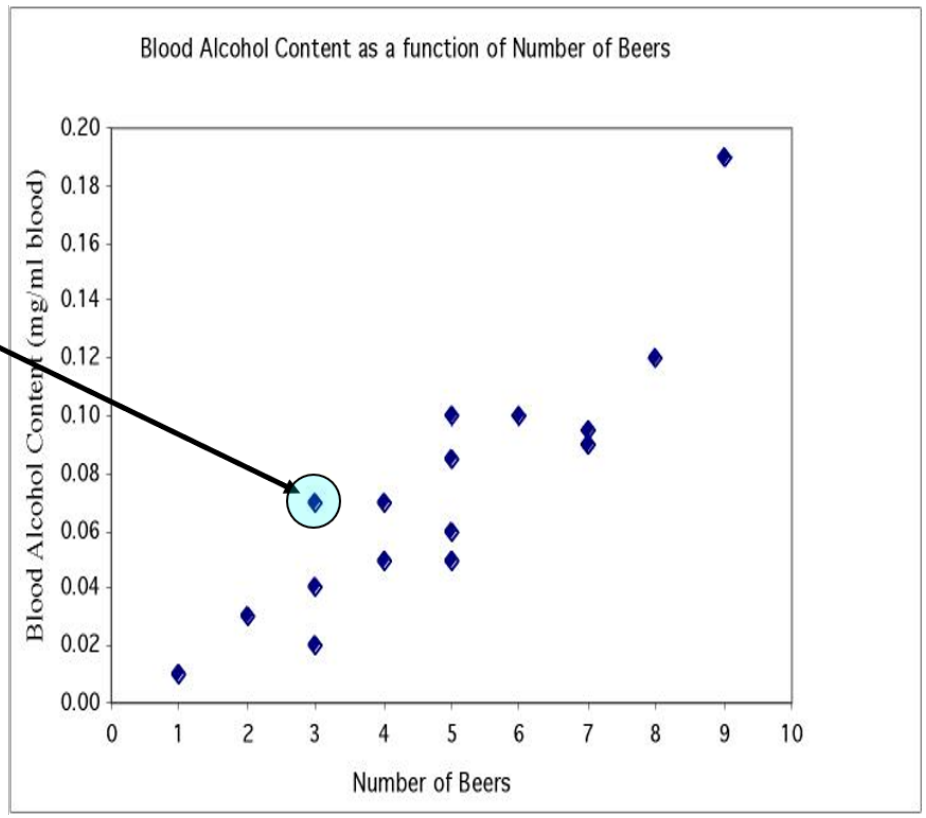
- Here, we have two quantitative variables for each of 16 students.
  1. How many beers they drank, and
  2. Their blood alcohol level (BAC)
- We are interested in the relationship between the two variables: How is one affected by changes in the other one?

Student	Beers	Blood Alcohol
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05

Scatterplots

In a **scatterplot**, one axis is used to represent each of the variables, and the data are plotted as points on the graph.

Student	Beers	BAC
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05

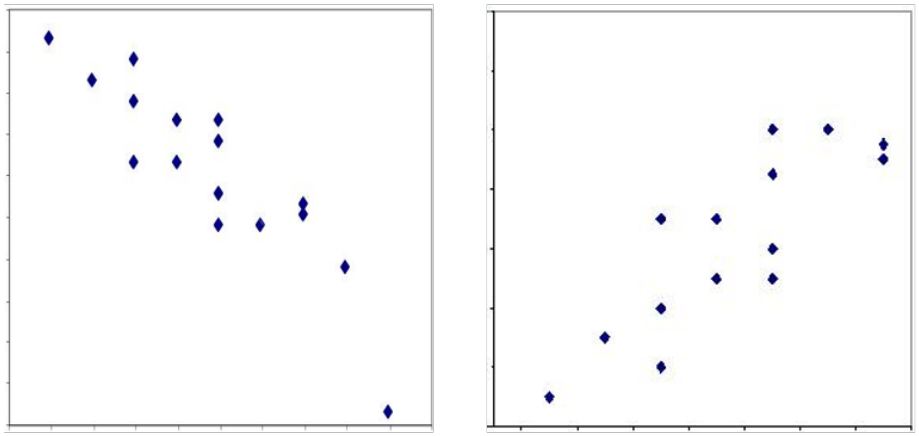


Interpreting scatterplots

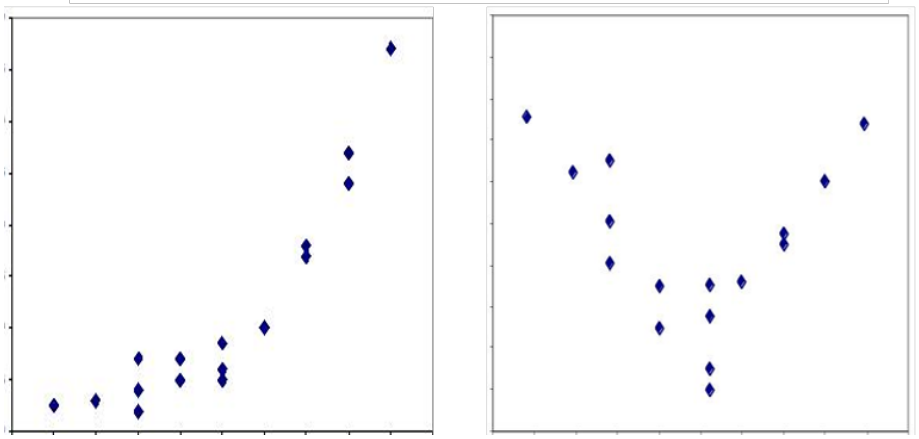
- After plotting two variables on a scatterplot, we describe the relationship by examining the **form**, **direction**, and **strength** of the association. We look for an overall pattern ...
  - Form: linear, curved, clusters, no pattern
  - Direction: positive, negative, no direction
  - Strength: how closely the points fit the "form"
- ... and deviations from that pattern.
  - Outliers

Form and Direction of an Association

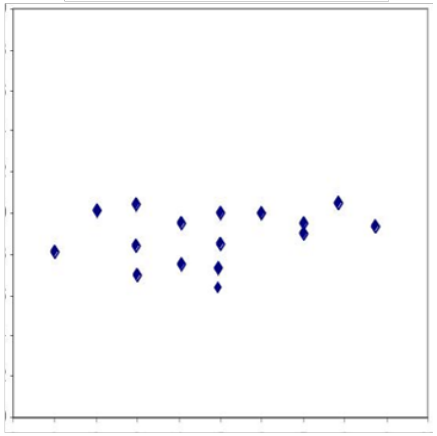
Linear

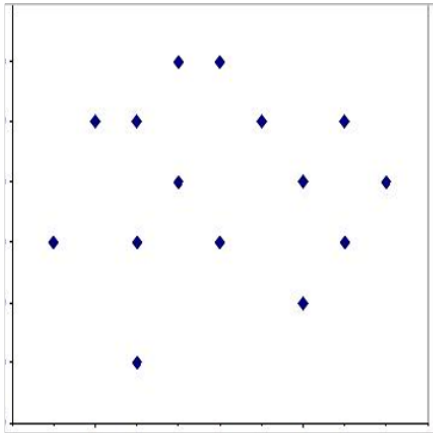


Nonlinear



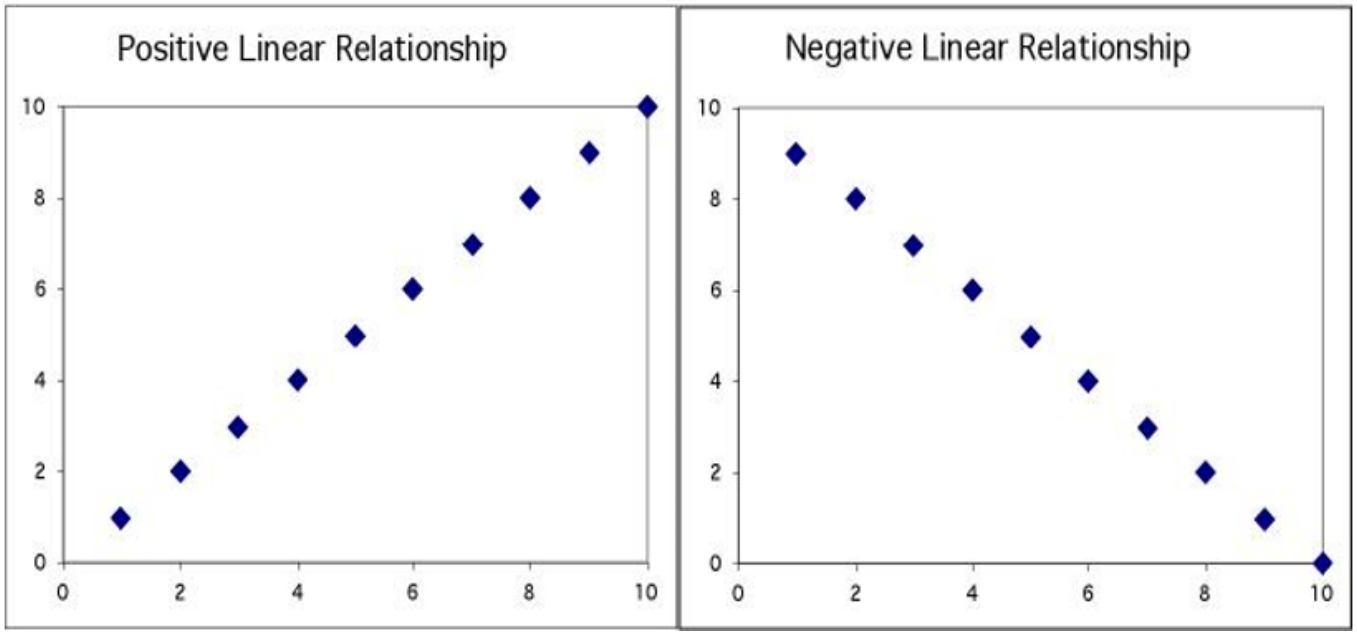
No relationship





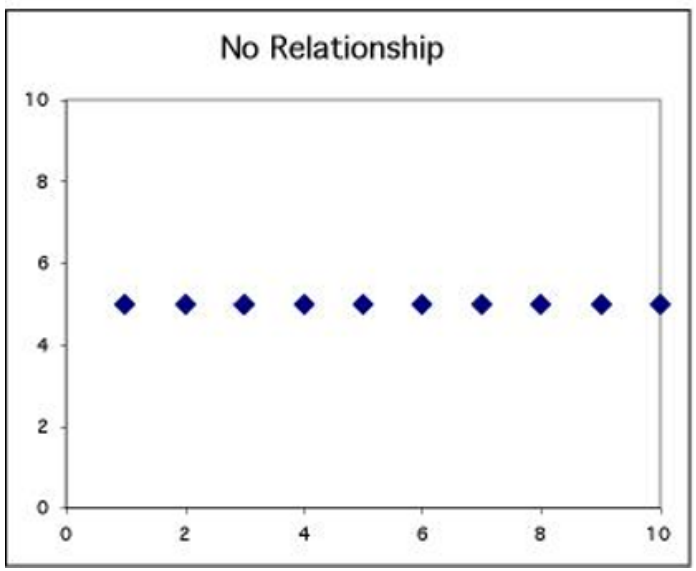
Direction of a Relationship

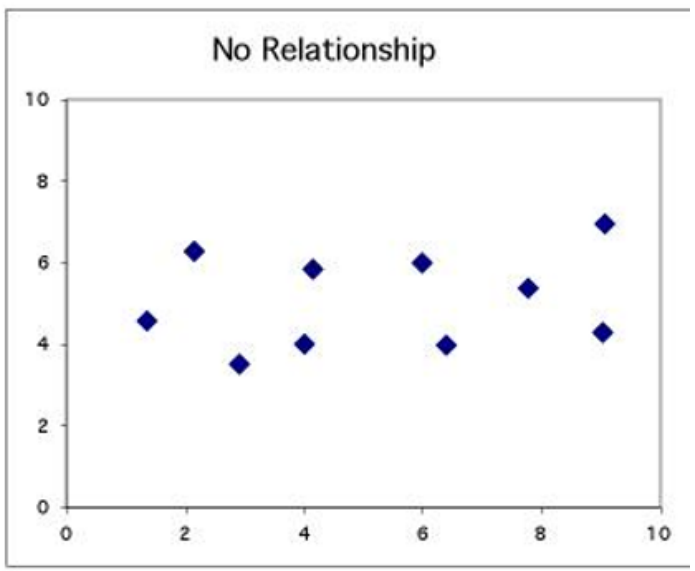
Positive association: High values of one variable tend to occur together with high values of the other variable.
Negative association: High values of one variable tend to occur together with low values of the other variable.



Direction of a Relationship

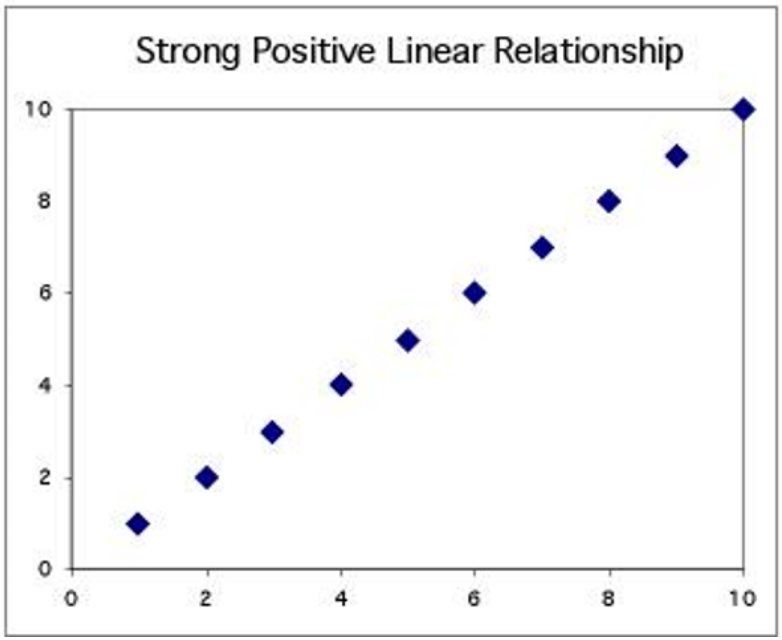
No relationship: X and Y vary independently. Knowing X tells you nothing about Y.



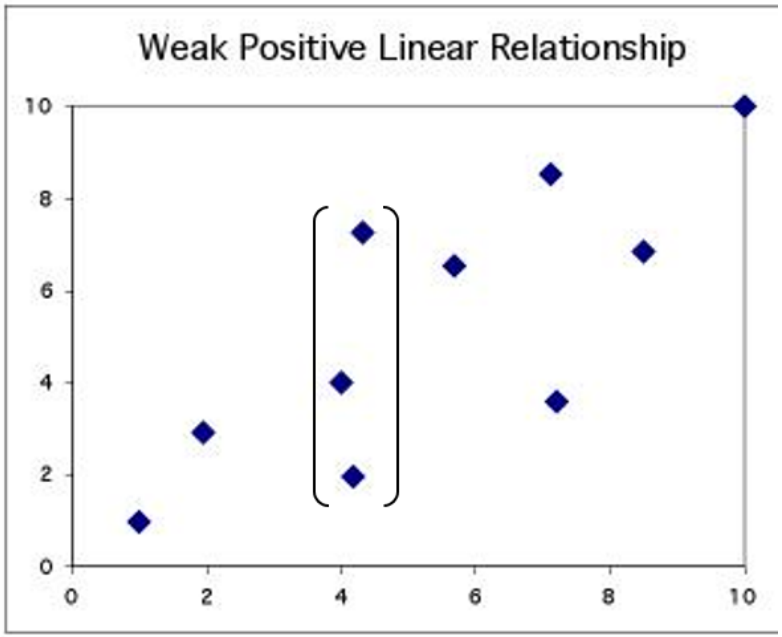


Strength of a Relationship

The **strength** of the relationship between the two variables can be seen by how much variation, or **scatter**, there is around the main form.

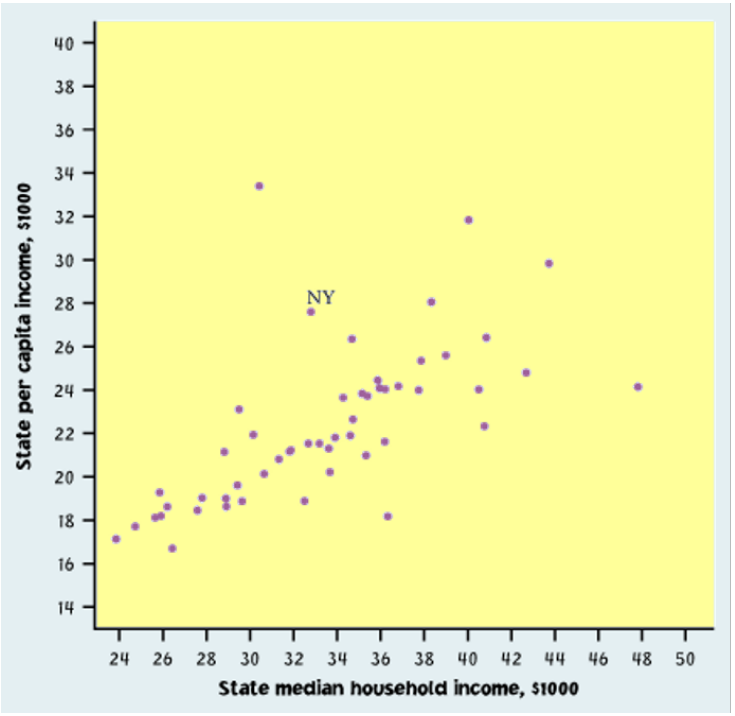


With a strong relationship, you can get a pretty good estimate of y if you know x.

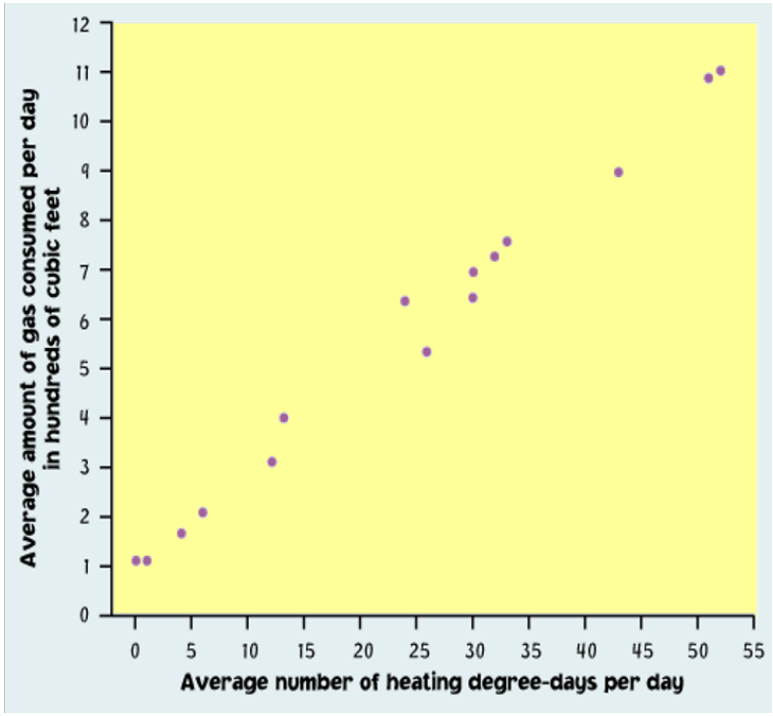


With a weak relationship, for any x you might get a wide range of y values.

Strength of a Relationship



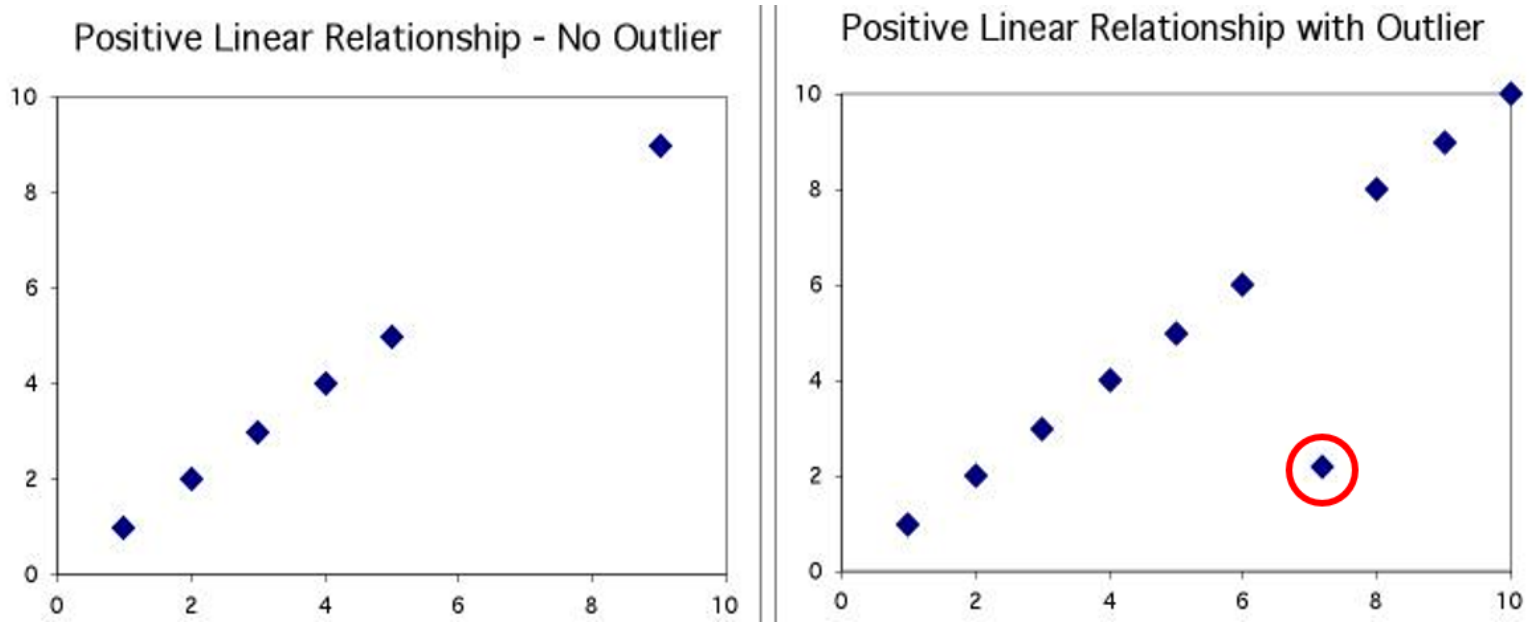
This is a **weak** relationship. For a particular state median household income, you can't predict the state per capita income very well.



This is a **very strong** relationship. The daily amount of gas consumed can be predicted quite accurately for a given temperature value.

Outliers

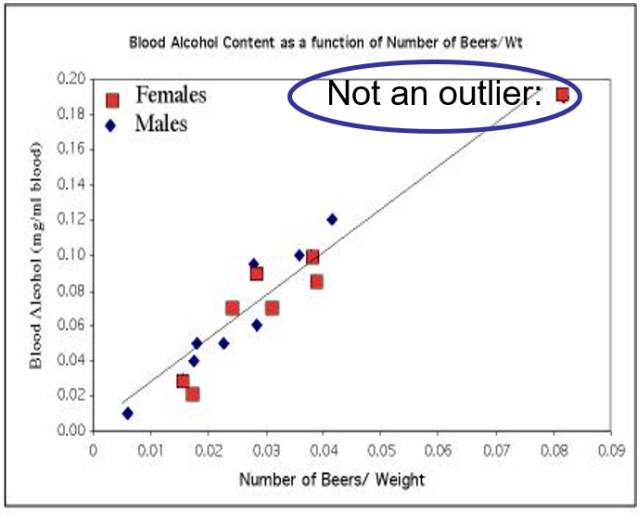
An outlier is a data value that has a very low probability of occurrence (i.e., it is unusual or unexpected).



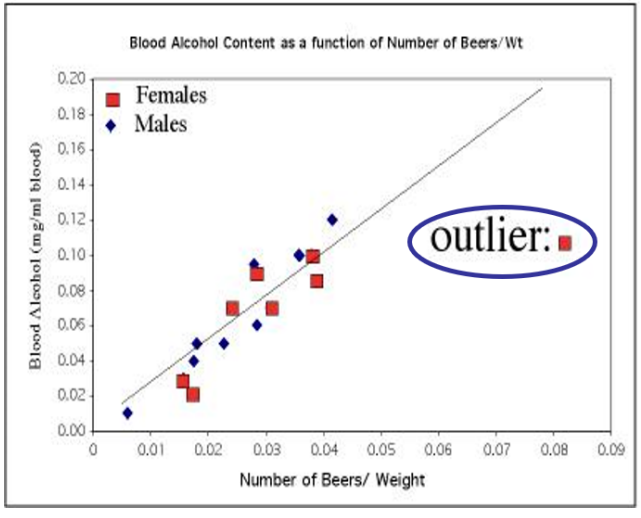
In a scatterplot, outliers are points that fall outside of the overall pattern of the relationship.

Outliers

- The upper right-hand point here is *not* an outlier of the relationship—It is what you would expect for this many beers given the linear relationship between beers/weight and blood alcohol.



- This point is not in line with the others, so it *is* an outlier of the relationship.



Measure of spread: the standard deviation

The standard deviation "s" is used to describe the variation around the mean. Like the *mean*, it is not resistant to skew or outliers.
Recall that there are two steps in the calculation of s: calculate the variance (s^2) and take the square root.

$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2} \tag{1}$$

The purpose of the standard deviation is to create a measure of spread that is *standardized*. For instance, the range of blood alcohol content values observed in a sample cannot be compared to the range of drinks consumed because they are measured with different units (parts of alcohol per 1000 parts of blood versus a count of beverages). Standard deviations use observations' distance from the average to create a measure of spread comparable across variables (i.e., a standard deviation increase has the same interpretation for both BAC and beers).

Correlation Coefficient: Pearson's "r"

The Pearson's "r" provides a way to more precisely measure the relationship between two variables with a measure that can be compared across relationships.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) \tag{2}$$

1. The steps to calculating the r coefficient begins with computing the mean and standard deviation of both variables you believe are related.
2. Then calculate the percent of a standard deviation each observation falls on both variables.
3. Multiplying these together for each variable provides a measure of the relationship between x and y for each observation in the sample.
4. Adding these factors and dividing them by the degrees of freedom $(n-1)$ provides the average strength of the relationship between x and y in the sample, or the r coefficient.

Example of Pearson's R: Airfare

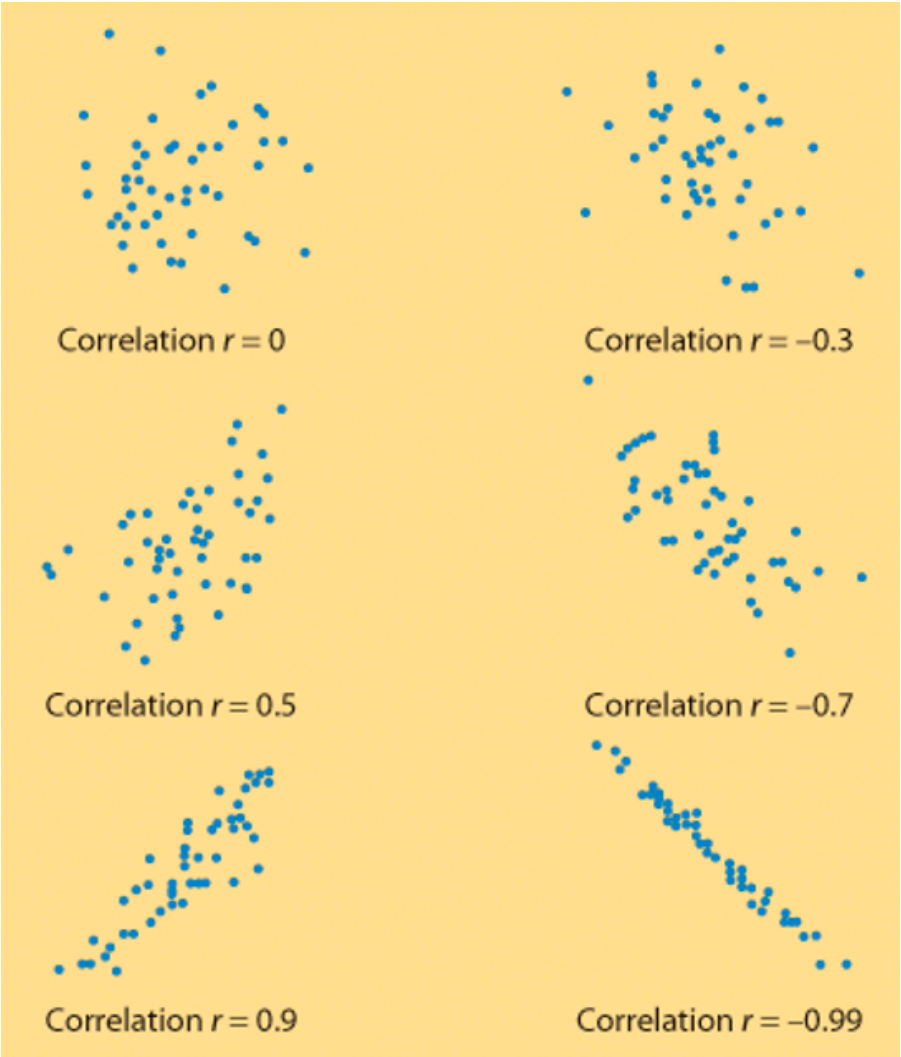
What's the relationship between the price of a plane ticket and the distance of the flight?

Sample: 15 flights
Avg. distance ($\overline{x}$): 1145 miles
s_x: 706.8 miles
Avg. price ($\overline{y}$): \$218.7
s_y: \$61.93
r ($\frac{\sum times}{n-1}$): 0.70
...a strong relationship!

idn	dist (x)	fare (y)	(x-xbar)/s	(y-ybar)/s	times
1	2310	361	1.65	2.30	3.79
2	656	132	-0.69	-1.40	0.97
3	904	274	-0.34	0.89	-0.30
4	444	182	-0.99	-0.59	0.59
5	2458	271	1.86	0.84	1.57
6	1050	153	-0.13	-1.06	0.14
7	1710	210	0.80	-0.14	-0.11
8	624	183	-0.74	-0.58	0.43
9	957	213	-0.27	-0.09	0.02
10	1334	167	0.27	-0.84	-0.22
11	444	186	-0.99	-0.53	0.52
12	769	203	-0.53	-0.25	0.14
13	810	253	-0.47	0.55	-0.26
14	453	190	-0.98	-0.46	0.45
15	2254	303	1.57	1.36	2.13

Visualizing r coefficients

- "r" quantifies the **strength** and **direction** of a linear relationship between 2 quantitative variables.
- **Strength**: how closely the points follow a straight line.
- **Direction**: is positive when individuals with higher X values tend to have higher values of Y.



Lurking Variables

A **lurking variable** is a variable not included in the study design that does have an effect on the variables studied.
Lurking variables can *falsely suggest* a relationship.
What is the lurking variable in these examples?
How could you answer if you didn't know anything about the topic?
Examples:

- Strong positive association between number of firefighters at a fire site and the amount of damage a fire does.
- Negative association between moderate amounts of wine drinking and death rates from heart disease in developed nations.