Statistical Methods-I STAT601
Lecture 1

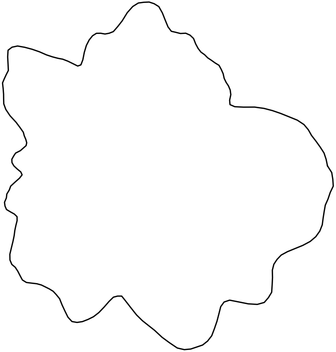
Coverage

- Univariate data
  - Graphical Tools :
    - Histogram
    - Boxplot
  - Numerical Tools :
    - Measures of locations: mean, median
    - Measures of dispersion: quartiles, range, IQR, SD
- Multivariate data
  - Scatter plot

Statistics

- Statistics is a discipline where relatively simple models are applied to approximately describe "random" phenomena observed in the real world and inference/prediction are made.
- Probability theory provides the mathematical foundations (17th and 18th centuries).
- The method of least squares was invented around the turn of the 19th century.
- Since then, many new techniques of probability and statistics have been or are being developed.
- Modern computers have expedited large-scale statistical computation, making new methods computationally feasible.

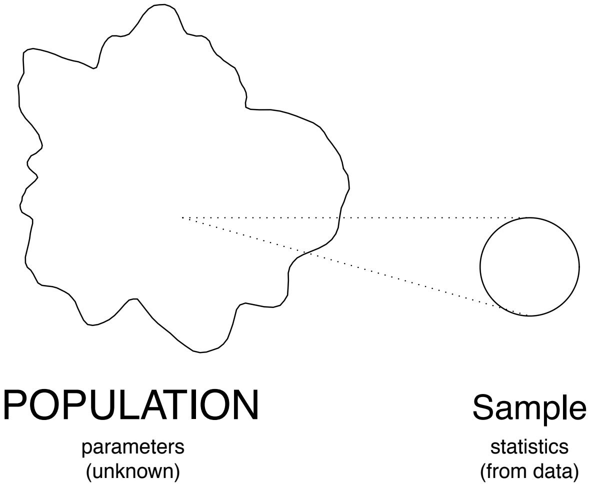
Probability vs. Statistics



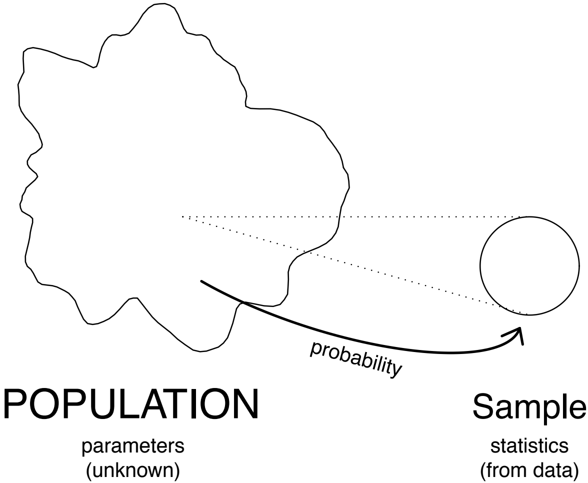
POPULATION

parameters
(unknown)

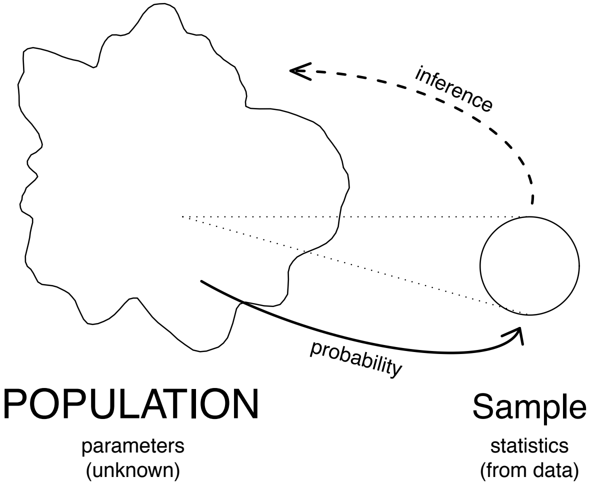
Probability vs. Statistics



Probability vs. Statistics



Probability vs. Statistics



Example: Iris data set of 5 variables for 150 observations

Index	Sepal-Length	Sepal-Width	Petal-Length	Petal-Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
...	...	...	...	...	
74	6.1	2.8	4.7	1.2	versicolor
75	6.4	2.9	4.3	1.3	versicolor
76	6.6	3.0	4.4	1.4	versicolor
...	...	...	...	...	...
...					
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

- R: str(...),head(...),tail(...)

Some Definitions

- **Unit/Subject/Individual**: each object or person in a population, sample, or experiment
- **Variable**: any characteristic of a unit
  - **Categorical (= Qualitative)**: places an individual into one of several groups or categories. **(Ordinal, Nominal)**
  - **Numerical (= Quantitative)**: taking numerical values on which we can do arithmetic.
    - **Discrete**: taking values from a discrete set of numbers (Ex: family size)
    - **Continuous**: taking values from a continuous set of numbers (Ex: time, weight, distance).
  - Ex., *Sepal.Length, Sepal-Width, Petal-Length* and *Petal-Width* are (continuous) numerical variables; *Species* is categorical variable.

Tabulate Data

- **Frequency table**, **Relative frequency table**
  Ex: `Species` and Sepal.Length of iris in the previous data

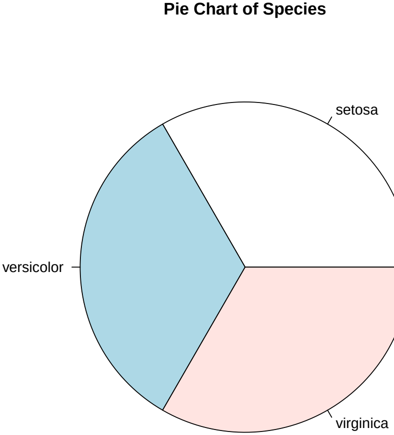
Species	Frequency	Relative Frequency
setosa	50	0.33
versicolor	50	0.33
virginica	50	0.33
Total	150	100%

S.Length	Frequency	Relative Frequency
(3,4]	0	0
(4,5]	32	0.213
(5,6]	57	0.380
(6,7]	49	0.327
(7,8]	12	0.080
Total	150	100%

- **Mode**: the most frequent observation
- R: table(), cut()

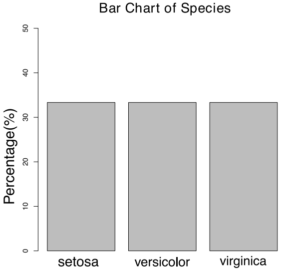
Categorical Data: Pie Chart

In a **pie chart**, the **area** of each slide is proportional to the percentage of individuals who fall into that category.



Categorical Data: Bar Graph/Barplot

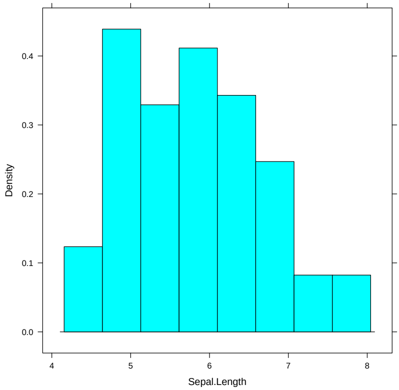
- The horizontal axis lists the categories
- X axis does not have a low end or a high end; The order of categories along the horizontal axis depends on the type of variable as well as the goal of the graph
- The height of the bars can be frequencies or percentages
- R: barplot(...)



Numerical Data: Univariate

- Graphical Tools:
  - Histogram
  - Boxplot
- Numerical Tools:
  - Measures of locations: mean, median, trimmed mean
  - Measures of dispersion: quartiles, range, interquartile range, standard deviation

What is Histogram?

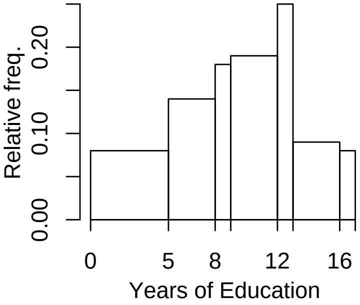


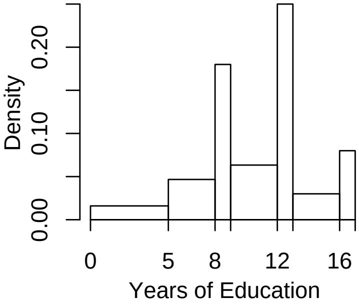
1. Group the observations into “bins” according to their value.
2. Count the individuals in each bin.
3. Draw the histogram:
   - R: histogram(...)
   - Label the horizontal axis with units of measurement.
   - How about the vertical axis?

Vertical axis: Frequency or Density?

The table on the right shows the years of educations for persons age 25 and over in the U.S. in 1960. The class intervals include the left endpoints, but not the right. Which histogram below makes more sense to you?

Years of education	Relative frequency
0 – 5	0.08
5 – 8	0.14
8 – 9	0.18
9 – 12	0.19
12 – 13	0.25
13 – 16	0.09
16 or more	0.08





Three common vertical scales of a histogram:

- **Frequency** of a class is the number of observations in that class (also called the "count").
- **Relative Frequency** $= \dfrac{\text{Frequency}}{\text{size of data set}}$ (also called the "proportion").
- **Density** $= \dfrac{\text{Relative Frequency}}{\text{Bar Width}}$

years of education	bin width	relative frequency	density
0 – 5	5	0.08	$0.08/5 = 0.0160$
5 – 8	3	0.14	$0.14/3 = 0.0467$
8 – 9	1	0.18	$0.18/1 = 0.1800$
9 – 12	3	0.19	$0.19/3 = 0.0633$
12 – 13	1	0.25	$0.25/1 = 0.2500$
13 – 16	3	0.09	$0.09/3 = 0.0425$
16 or more	1	0.08	$0.08/1 = 0.0800$
Total		1.00	

Heights of Bars

- If set the bars with **unequal width**, then the **area**, not the **height**, of each bar is proportional to the frequency of that class.

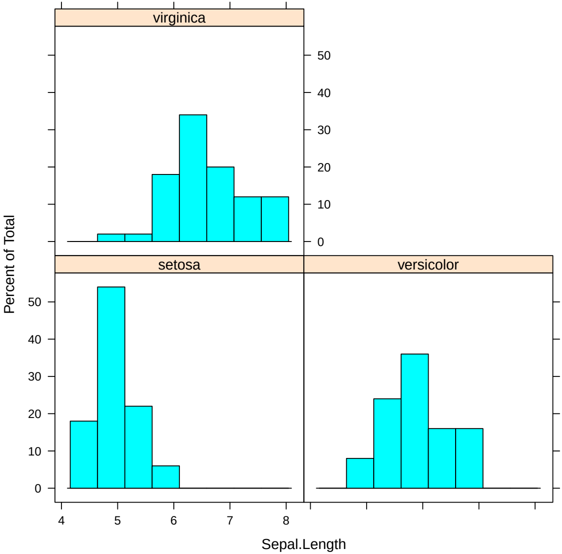
$$\begin{gathered} \mathit{area} \propto \ \mathit{frequency} \\ \mathit{bar\ height} \ \propto \ \mathit{density} = \frac{\mathit{frequency}}{\mathit{bar\ width}} \end{gathered}$$

- If set the bars with **equal width**, then the height of each bin is just proportional to it's frequency.

$$\mathit{bar\ height} \ \propto \ \mathit{frequency}$$

- With equal width, the height does not have to be the density, but can be the frequency (or relative frequency).

What to Look in a Histogram?

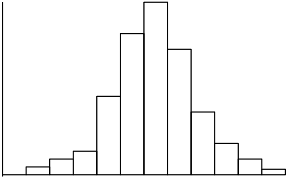


What to Look in a Histogram?

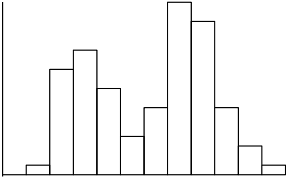
- **Shape**: mode and skewness of a histogram
- **Center**: Where is the "middle" of the histogram?
- **Spread**: What are the smallest and largest values?
- **Outliers**: Are there any observations that lie outside the overall pattern? They could be unusual observations, or they could be mistakes. Check them!

Mode of Histograms (= Number of Peaks)

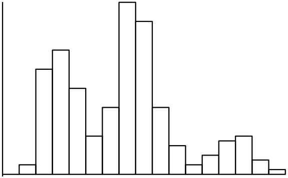
Unimodal



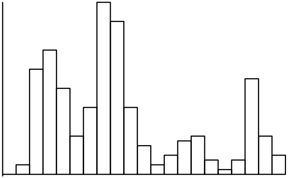
Bimodal



Trimodal

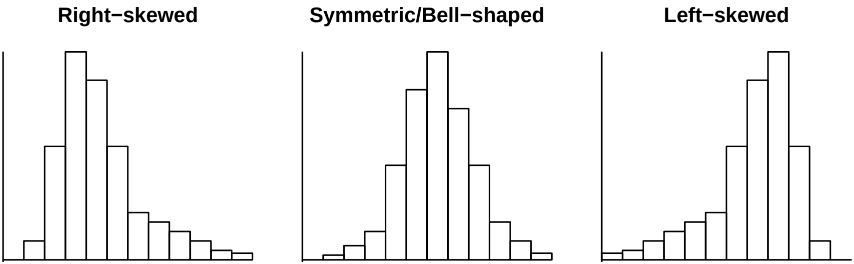


Multimodal



A histogram with two or more modes may indicate that the data is a mixture of two or more distinct populations.

Skewness of Histograms



Mean

The **mean** of a set of observations is the arithmetic average of the observations:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ex: Say the age of 9 individuals are

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
43	35	43	33	38	53	64	27	34

The mean age of the 9 people is given by:

$$\bar{x} = \frac{43 + 35 + 43 + 33 + 38 + 53 + 64 + 27 + 34}{9} = \frac{370}{9} = 41.11.$$

Median

For a list of numbers, the **median** is a number such that half of the list are smaller than it and half of the list are larger than it.

How to find the median of a list of numbers $x_1, x_2, \dots, x_n$?

1. Sort the list from the smallest to the largest:
$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}.$$
2. If there are odd numbers in the list (n is odd), the middle number in the sorted list is the median:
$$Median = x_{((n+1)/2)}$$
3. If there are even numbers in the list (n is even), the average of the two middle numbers in the sorted list is the median:
$$Median = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Example 1: Say the age of 9 individuals are

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
43	35	43	33	38	53	64	27	34

The median is ____________

Example 2: Say the age of 10 individuals are

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
43	35	43	33	38	53	64	27	34	27

The median is ____________

Robustness of the Median

Consider the list -2, -1, 0, 0, 2, 4. If the number '2' in the list is miss recorded as 20,

- The mean is increased by $(20 - 2)/6 = 3$.
- The median is unaffected.

Median is more resistent, i.e., less sensitive to extreme values or outliers than the mean. We say the median is more **robust**.

- Example: Housing sales price in Hyde Park (Jun - Aug, 2011)
  Mean :$525,384, Median: $227,000[1].

[1] Source: http://www.trulia.com/home prices/Illinois/Chicago-heat map/

Measure of Spread: Percentiles

- Sample median is a special example of a sample quantile (or percentile).
- Denote the pth sample quantile as $y_{[p]}$ with $0 < p < 1$.
- To compute the pth sample quantile:
  1. Arrange data in a list of ascending order.
  2. Compute $n \times p$.
  3. If $n \times p$ is an integer, then $y_{[p]}$ is the average of $(n \times p)$th and $(n \times p + 1)$th data values in the list.
  4. If $n \times p$ is not an integer, then round up to $\lceil n \times p \rceil$ and use the $\lceil n \times p \rceil$th data value in the list.

Example

For the list below

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
2	8	3	4	6	5	4	1	5	3	2	7	11	1

Arrange in increasing order:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$
1	1	2	2	3	3	4	4	5	5	6	7	8	11

- $p = 0.10$: $np = 1.4 \implies$ 10th percentile is $x_{(2)} = 1$
- $p = 0.25$: $np = 3.5 \implies$ 25th percentile is $x_{(4)} = 2$
- $p = 0.50$: $np = 7.0 \implies$ 50th percentile is $\frac{x_{(7)}+x_{(8)}}{2} = 4$
- $p = 0.75$: $np = 10.5 \implies$ 75th percentile is $x_{(11)} = 6$
- $p = 0.80$: $np = 11.2 \implies$ 80th percentile is $x_{(12)} = 7$

Quartiles, IQR, Range, Five-Number Summary

- The 25^{th}, 50^{th}, and 75^{th} percentiles are called **quartiles**:

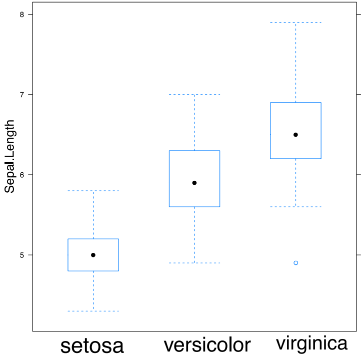
$$25^{\text{th}}\text{percentile} = \text{first quartile}(Q_1)$$
$$50^{\text{th}}\text{percentile} = \text{second quartile}(Q_2) = \text{the median}$$
$$75^{\text{th}}\text{percentile} = \text{third quartile}(Q_3)$$

- The **Interquartile Range (IQR)** $= Q_3 - Q_1$
  - One of the measures of spread
  - Not sensitive to extreme values
- **Range** $= \text{max} - \text{min} = x_{(n)} - x_{(1)}$
- **Five-Number Summary**:

$$x_{(1)} \quad Q_1 \quad \text{Median} \quad Q_3 \quad x_{(n)}$$

Boxplot

A boxplot is a graph of the five-number summary.



- Quartiles: Bottom and Top of the box
- Median: A dot in the box
- Lines: from the ends of the box to the most extreme observations within a distance of 1.5 IQR (Interquartile range).
- Outliers: outside 1.5 IQR from the ends of the box.

Variance and Standard Deviation

Suppose there are n observations $x_1, x_2, \ldots, x_n$.
The **variance** of the n observations is:

$$\begin{aligned} s^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1} \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

This is (approximately) the average of the squared distances of the observations from the mean.
The **standard deviation** (SD) is:

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

Why $n-1$?

Division by $n-1$ instead of n in the variance calculation is a common cause of confusion. Why $n-1$? Note that

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

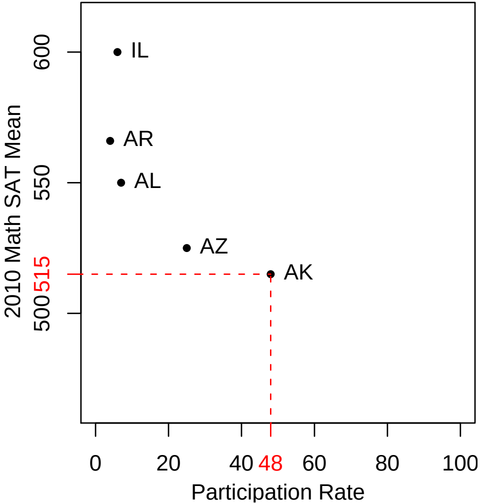
Thus, if you know any $n-1$ of the differences, the last difference can be determined from the others. The number of "freely varying" differences, $n-1$ in this case, is called the **degrees of freedom**.

A better reason that will be explained in later lectures is that dividing by $n-1$ makes the sample variance *unbiased.*

Multivariate Data

2010 Mean Math SAT Scores[a]

State	Participation Rate[b]	Math
AK	48	515
AL	7	550
AR	4	566
AZ	25	525
CA	50	516
CO	18	572
	⋮	
IL	6	600
	⋮	
WY	5	567



[a]Source: CollegeBoard

[b]The percentage of high school graduates in the class of 2010 who took the SAT

Scatter Plot

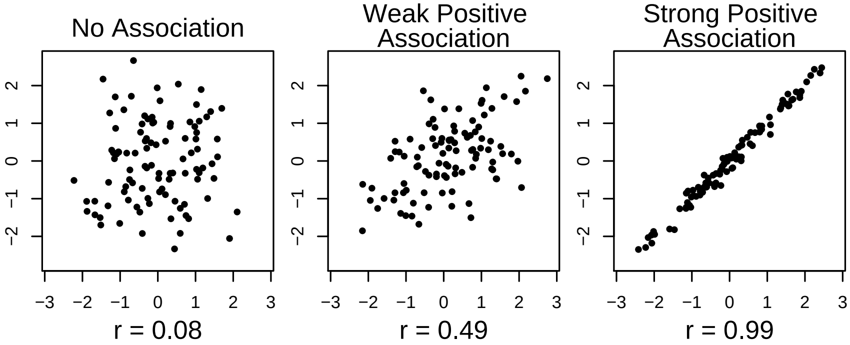
A **scatter plot** shows the relationship between two numerical variables measured on the same units.
The values of one variable are on the x-axis, and the values of the other are on the y-axis. Each individual is represented by a point in the graph.

What to Look in a Scatter Plot?

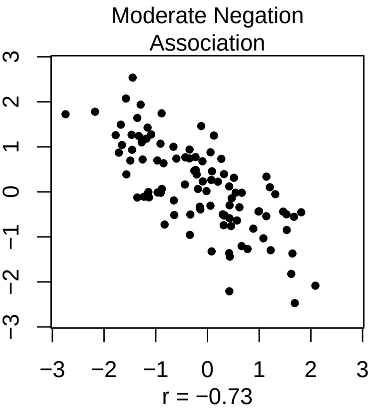
1. What is the overall pattern?
   - What is the *form* of the relationship?
     (linear, curved, clustered ...)
   - What is the *direction* of the relationship?
     (positive association, negative association)
   - What is the *strength* of the relationship?
     (strong, weak, ...)

2. Are there any deviations from the overall pattern? An individual that falls outside the overall pattern is an **outlier**.

Strength of Association

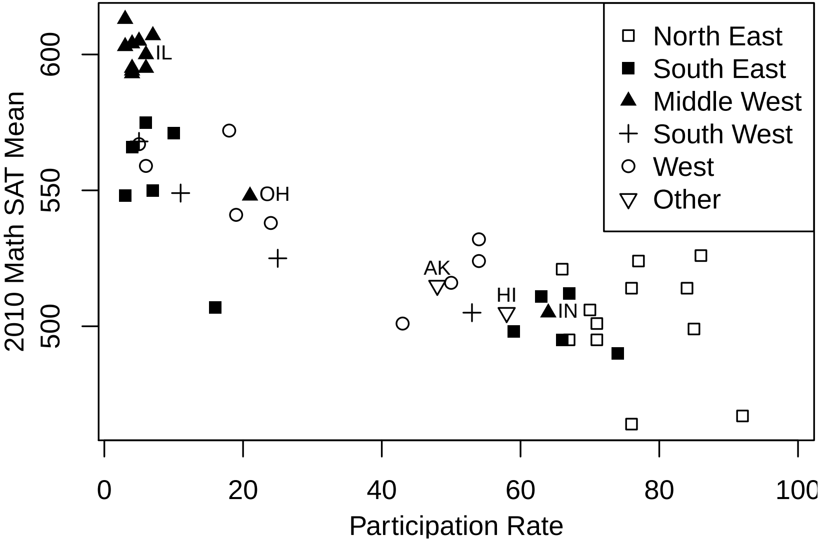


Negative Association

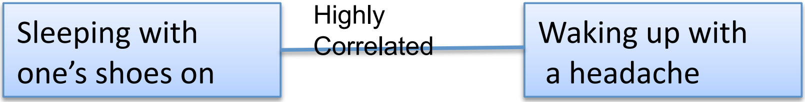


Adding Categorical Variables to Scatter Plots

Points in different categories can be marked with different colors or symbols.



Correlation $\neq$ Causation



Correlation $\neq$ Causation