Sampling and Summary Statistics

Biology 683

Lecture 2

Heath Blackmon

Last week

- What are some causes of the reproducibility crisis?
- What makes R awesome?

Today

1. Terminology
2. Summarizing Data
3. Central Limit Theorem
4. Covariance and Correlation

Populations and Samples

- **Populations**

  Some sort of group of something - could be anything
  - Undergraduates at Texas A&M
  - Jewel beetles in Arizona
  - Strain of flies in the lab

- **Samples**
  - A subset of individuals drawn from a population

What is the population?

We wanted to examine any association between the severity of injuries, and the height from which cats fall in high-rise buildings.

In the period between January 1, 1998 and December 12, 2001 at the Clinic of Surgery, Orthopaedics and Ophthalmology of the Veterinary Faculty, 119 cats were treated after a fall or jump from a balcony or window, where the owners saw the fall, or where there was a reasonable suspicion that a fall had occurred. Only those cats that fell from the second or higher stories were included. The owners brought the cats for treatment within varying periods of time after the fall (from 30 min to over a month).

Vnuk, et al. "Feline high-rise syndrome: 119 cases (1998–2001). *Journal of Feline Medicine & Surgery* 6.5 (2004): 305-312.

Sampling Considerations

Target population

- Need to sample a representative population
- A sample of people from College Station, for instance, would probably not be representative of New Yorkers

Sampling Error

- Chance alone will cause your sample to depart from the population

Parameter, estimates, sampling considerations

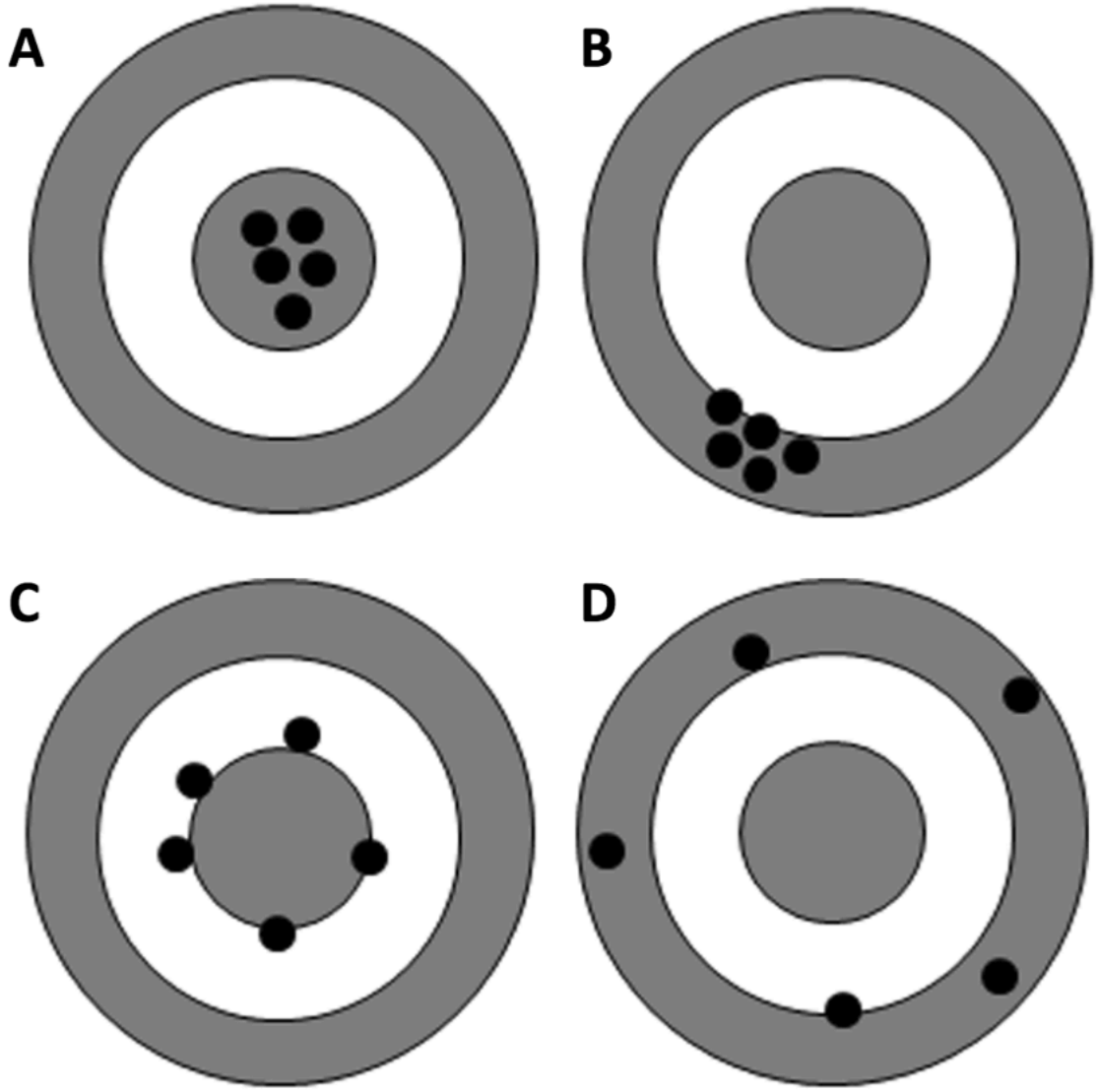
Parameter: Population-level variables we are trying to estimate

Estimate or Statistic: The value of the parameter inferred from the sample

Bias: If something about the sampling procedure causes the sample to systematically misrepresent the population.

Precision: How tightly grouped are the estimates?

Accuracy vs Precision



- Precision is a measure of spread
- Accuracy is a measure of bias

Random Sampling

1. Every unit in a population should have an equal chance of being sampled.

2. The selection of units must be independent.

3. Lots of ways of being non-random…

Your big idea should be a hypothesis

A statistical hypothesis is a specific claim about a population parameter

Caloric restriction increases the lifespan of Drosophila melanogaster.

The rate of evolution in wingless species is higher than winged species.

Pesticide exposure causes feminization of amphibian males.

Repetitive DNA content is higher in venomous than nonvenomous reptiles

Data

Variables
The characteristics that differ among individuals

Data
The measurements of variables taken for a sample of individuals

Categorical Variables
Individuals are in qualitative categories

Data

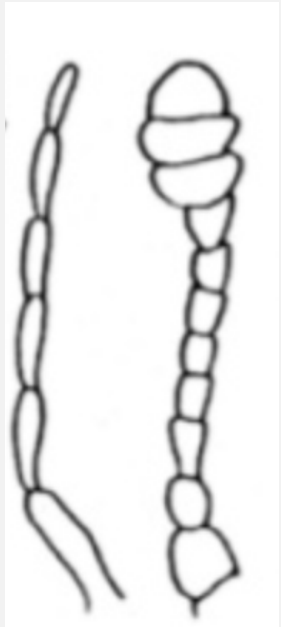
Numerical Variables
Individuals vary on a quantitative scale

Ordinal
The categories can be ordered

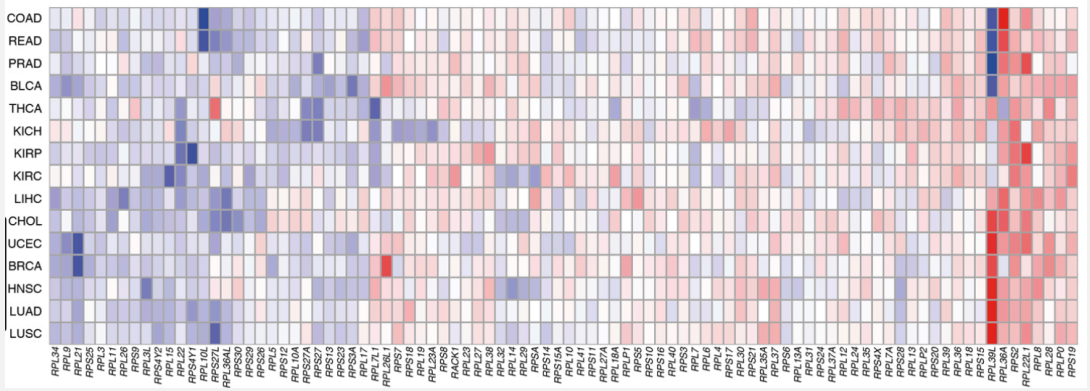
Nominal
The categories have no inherent order

Continuous vs Discrete

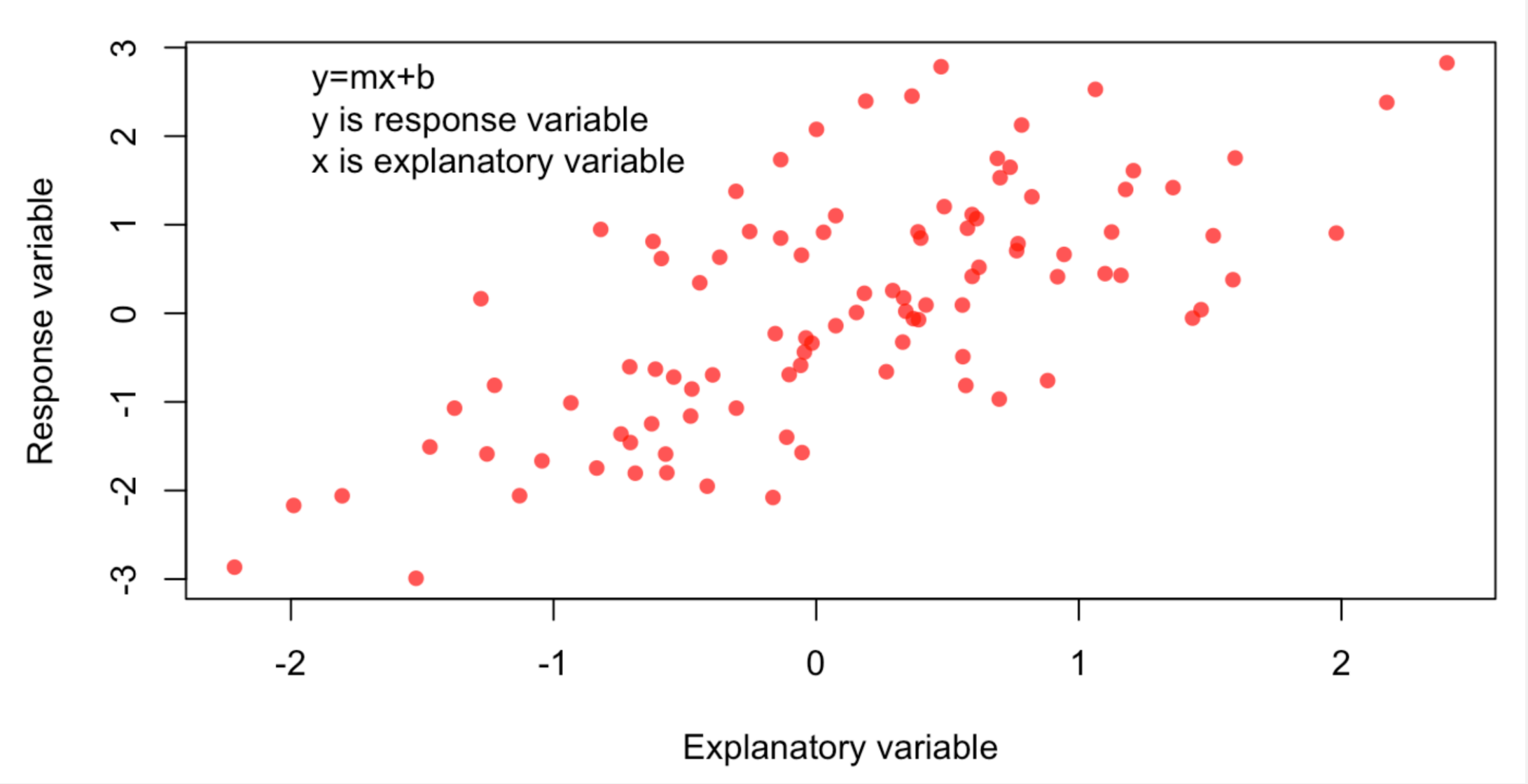
Continuous variables
a variable that has an infinite number of possible values



Discrete variables
a variable that has a finite number of possible values



Explanatory and Response Variables



Experimental vs observational studies

- Does caloric restriction increase lifespan in mice?
- Is global warming caused by human activities?
- Does smoking cause lung cancer in humans?
- Does parasite infection reduce mating success of beetles?
- Does oxytocin affect sexual attraction in humans?
- Do sex chromosomes increase the rate of speciation?

Summarizing data is necessary and prefered

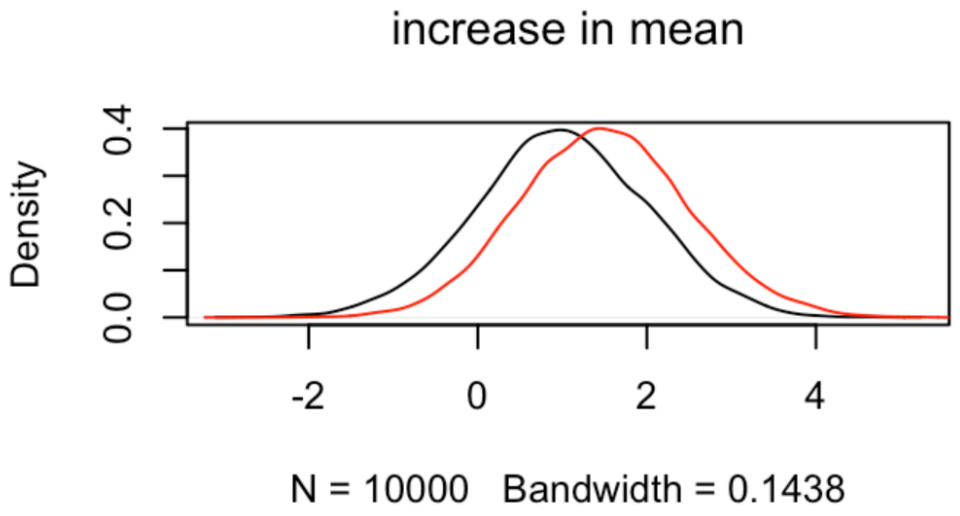
- Many datasets are simply too big to look at all values and form an impression?

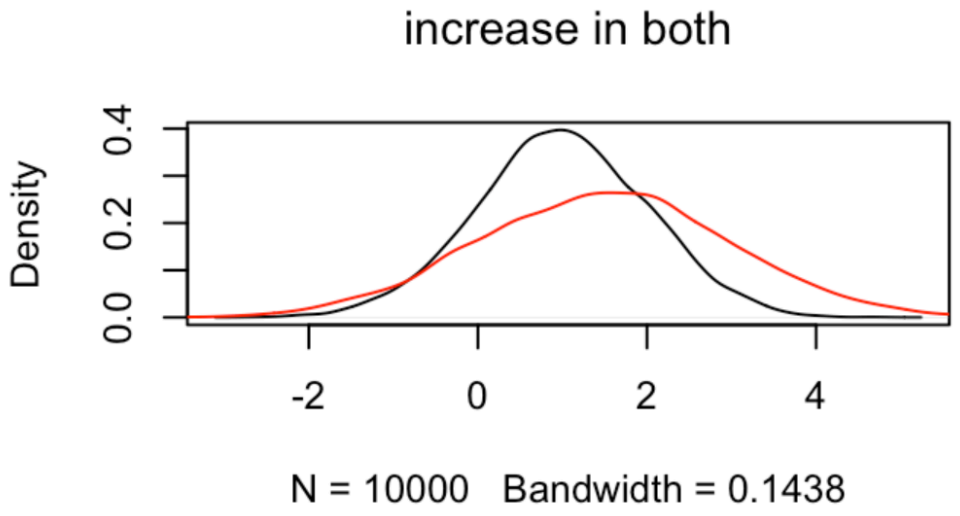
- Our impressions of small datasets are often misled by our tendency to look for patterns.

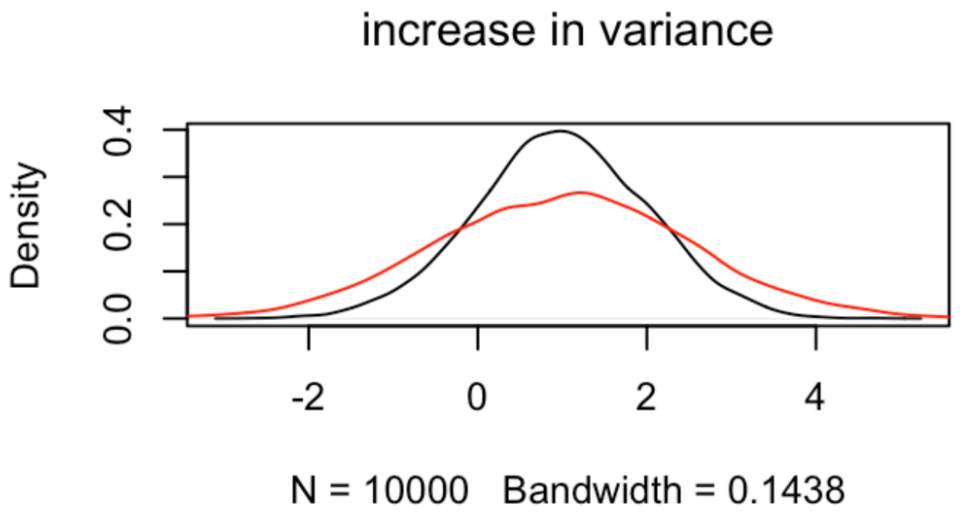
Typical summary statistics

- **Mean:** Sum of the observations divided by the number of observations
- **Median:** The middle observation in a set of data
- **Variance:** The average squared deviation from the mean
- **Standard Deviation:** The square root of the variance

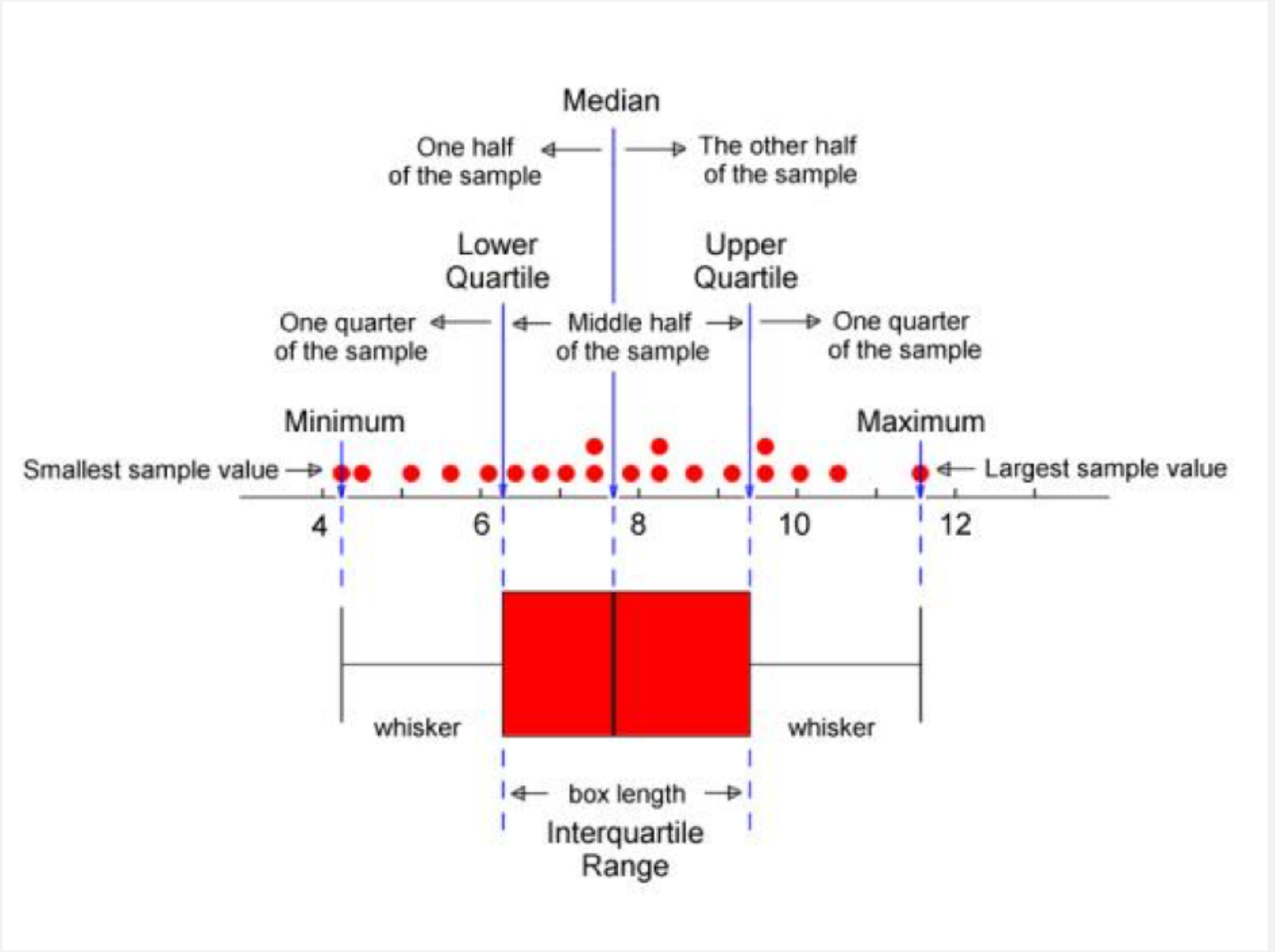
Mean and variance







Box Plot


Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
Middle half of the sample
One quarter of the sample
Minimum
Maximum
Smallest sample value
Largest sample value
4
6
8
10
12
whisker
whisker
box length
Interquartile Range

Estimating with uncertainty

Samples versus Populations

The mean or standard deviation statistic you calculate from your sample is an estimate of the population parameter.

Parameter Symbols:

μ : population mean

σ : population standard deviation

Statistic Symbols:

$\bar{Y}$: sample mean

s : samples standard deviation

For a sample of a population

The mean is just: $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

The standard deviation is $s = \sqrt{s^2}$

Where s^2 or the variance is: $s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$

Central limit theorem

- Imagine that we sample from the same population many times, so we have a bunch of different, independent samples.

- Each sample will have a mean, but the means will be different due to chance In principle, we could draw a histogram of these means.

- In general, you only have one sample from a given population, however, so what can you infer about the distribution of the means from your sample?

- The Central Limit Theorem states that regardless of the underlying population distribution of the variable of interest, the distribution of the population of means will be roughly normal.

Central limit theorem

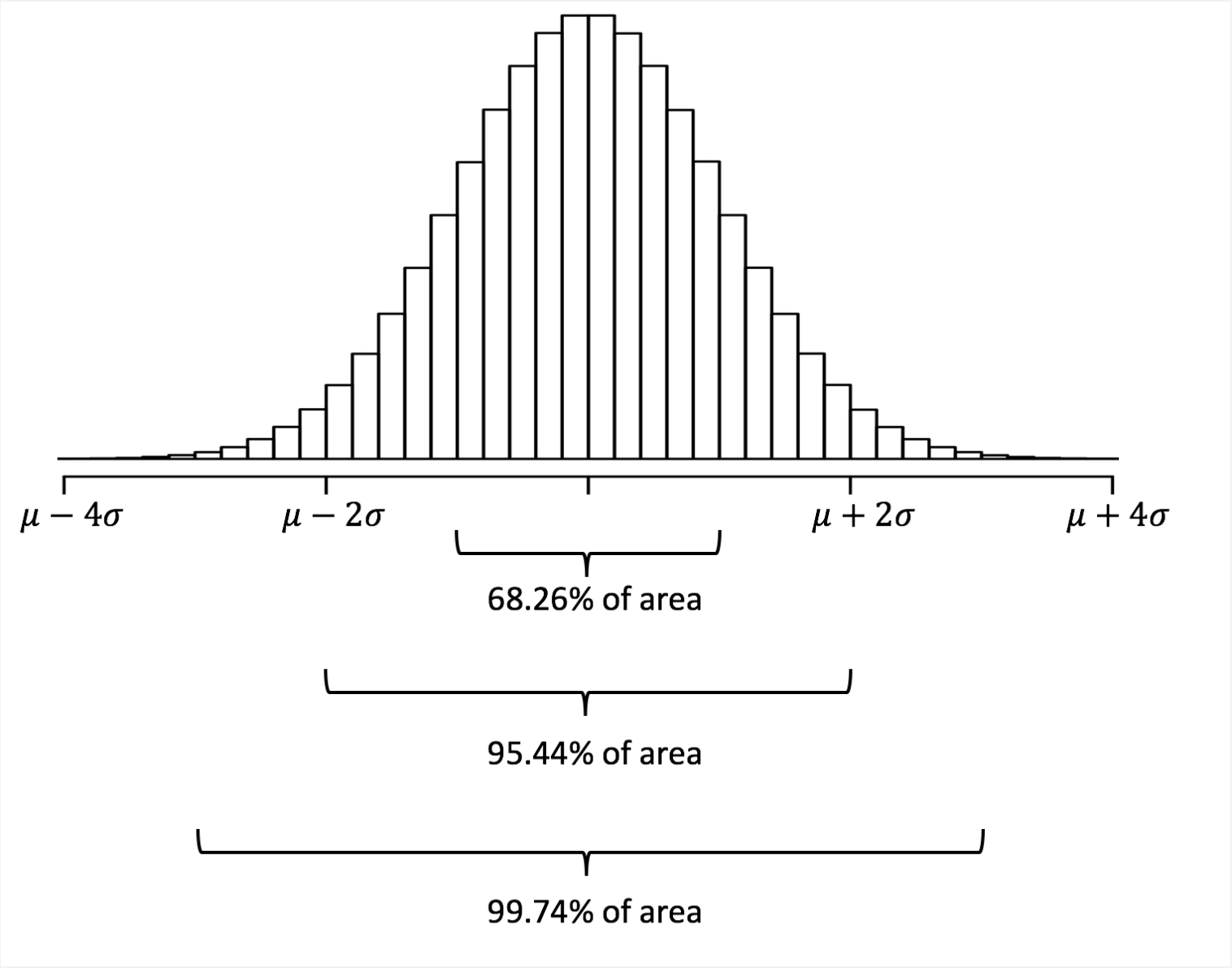
Your estimate of the sample mean is an estimate of the mean of this distribution of means (that is, it's your best estimate of the population mean).

The hypothetical distribution of sample means has a standard deviation equal to s divided by the square root of n.

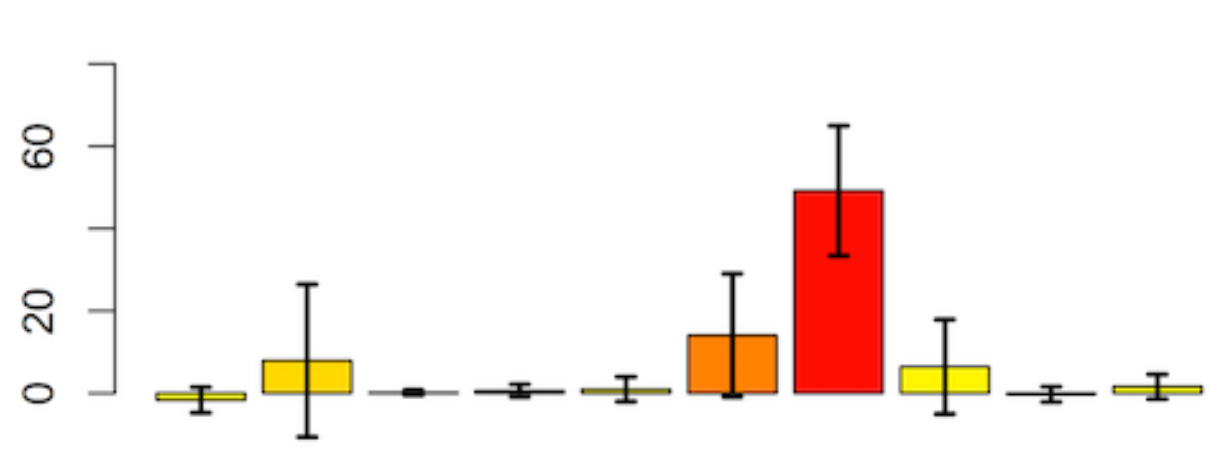
$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

We call this standard deviation the standard error of the mean.

Estimating with uncertainty

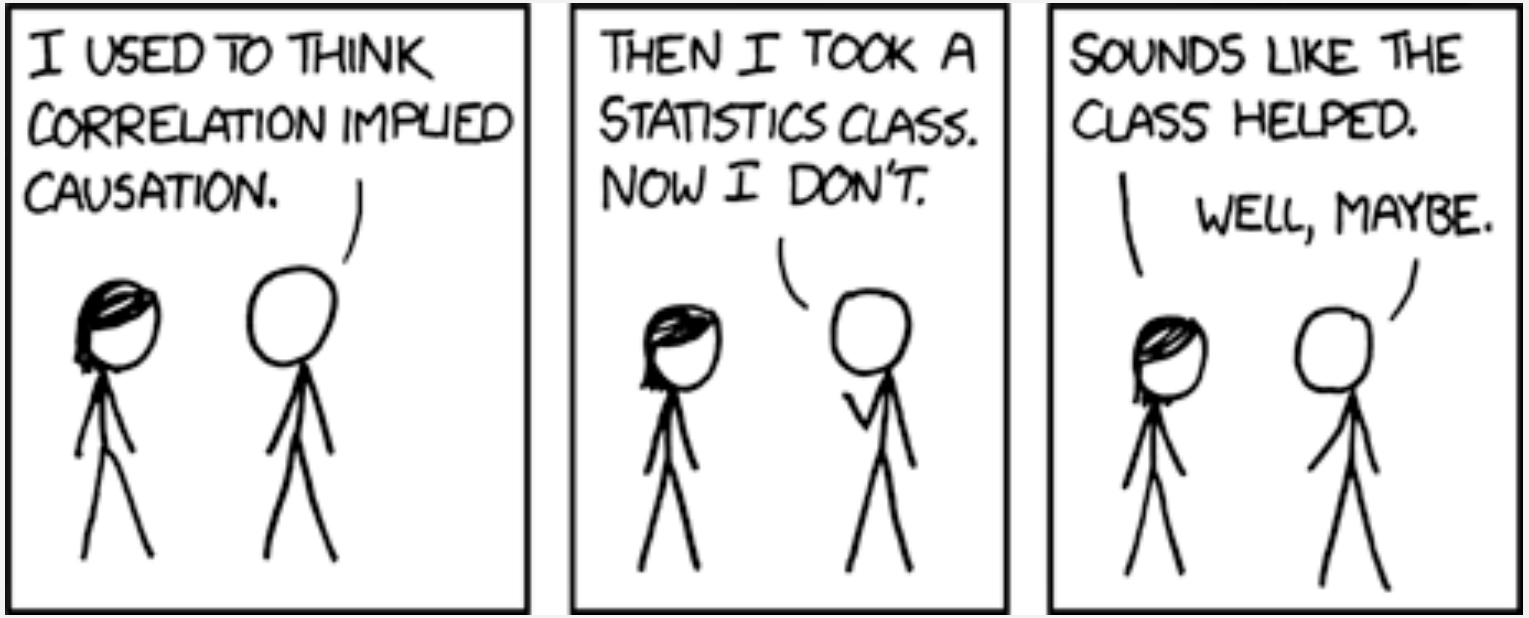


Error bars

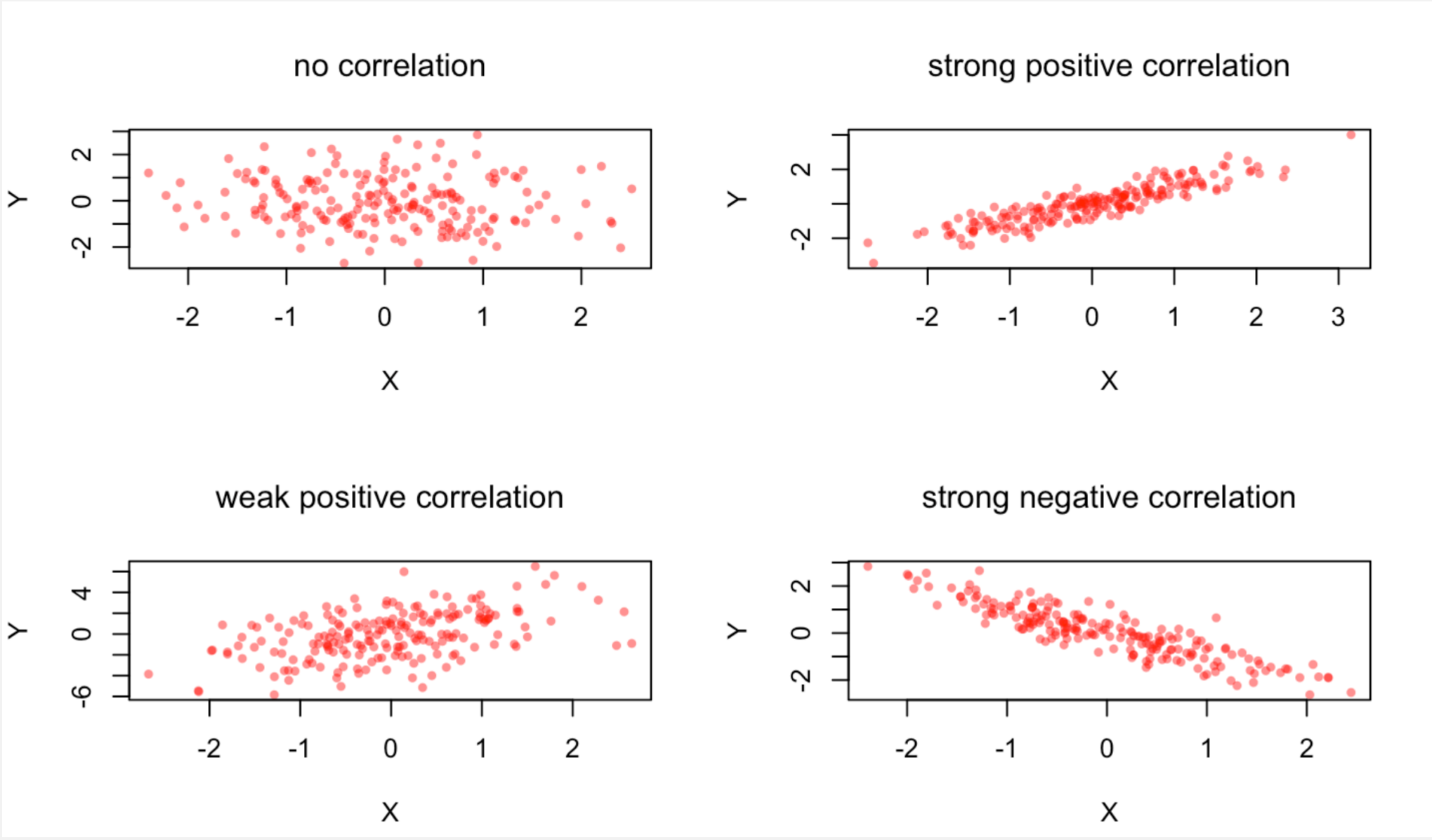


- Error bars can be a useful way to show uncertainty when it's not possible to show the actual data points.
- Usually, they represent 1 SE or the 95% CI, but not always.
- **THE FIGURE LEGEND SHOULD INDICATE WHAT THE ERROR BARS REPRESENT!**

Covariance and Correlation



Covariance and Correlation



Covariance and Correlation

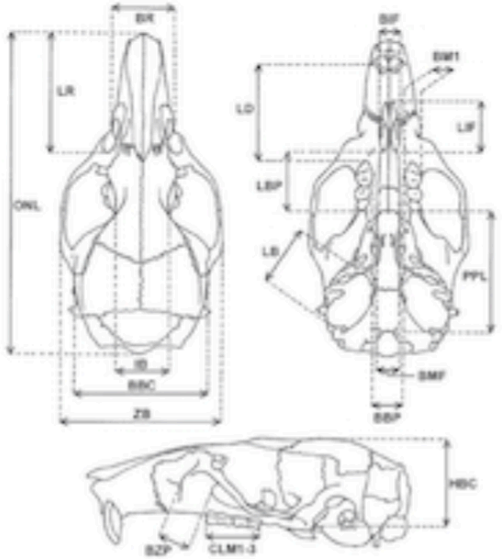
The covariance shows the extent to which the two variables are not statistically independent

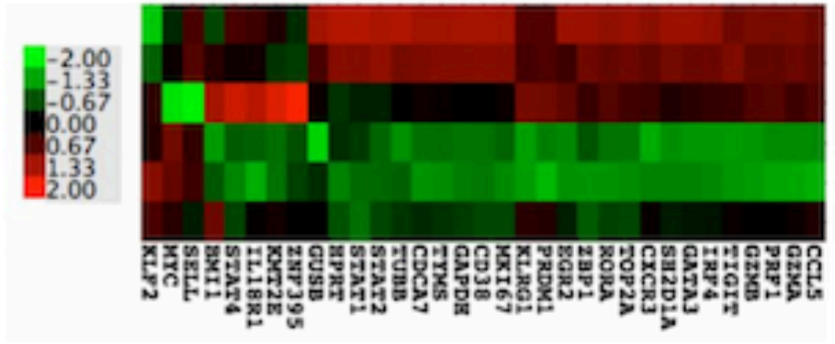
$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

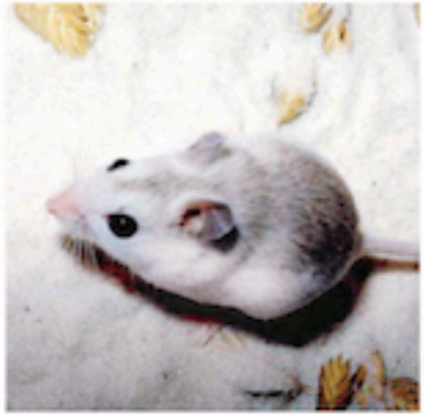
The correlation is the covariance, standardized to fall between -1 and 1.

$$r(X,Y) = \frac{cov(X,Y)}{s_x s_y}$$

Biological examples of correlations







Your turn

- Lets demonstrate that the means of samples from an exponential distribution are normally distributed.

  You will need:
  `rexp, hist, mean`

  Might use:
  `for, sample`

For Thursday

1. Read chapters 3 and 4 of WS
2. Install R and Rstudio on a laptop

Bring laptop to class!

Heath Blackmon
BSBW 309A
coleoguy@gmail.com