Statistics Bootcamp Day 2

17 September 2019

IRiSS

Welcome to bootcamp!

Our goals:

1. Increase students’ understanding of and confidence with basic statistical concepts.
2. Build students’ programming intuition and data management skills.
3. Encourage collaboration and camaraderie among the graduate student cohort.

Overview of the week

~~Monday: mindset, descriptive & inferential statistics, summary statistics, and Stata workshop~~

Tuesday: graphing, exponents/logarithms, sampling distributions, and statistical significance

Wednesday: probability basics, file structure and data workflow

Thursday: variable types, functions, lines of best fit, prediction equations

Friday: matrix algebra basics, reading calculus

Today’s learning objectives

...conduct basic exponent and logarithm computations

...present data as graphs and tables

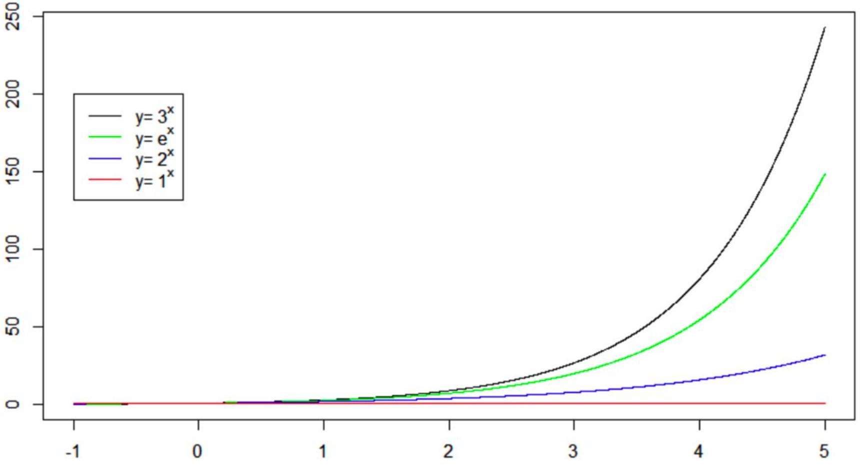
...explain the difference between a population distribution, sample distribution, and a sampling distribution

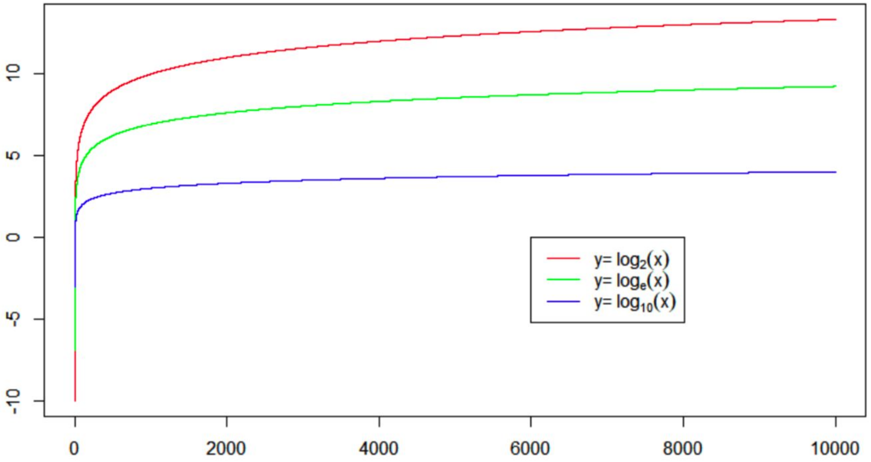
...explain the logic of statistical significance and repeated sampling

Exponents and Logarithms

Exponential Functions

Logarithmic Functions





$$y = a^x$$

$$y = \log_a(x)$$

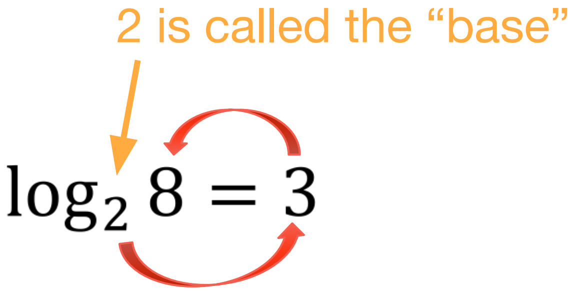
$2 * 2 * 2 = 2^3 = 8$

$\log_2 8 = 3$

Logarithms

- The inverse function of exponents:

$$2^3 = 8$$



2 is called the “base”

$$\log_2 8 = 3$$

- If no base is written, it is an implied base 10.

$\log 1000$ is the same as $\log_{10} 1000 \;\; = 3$

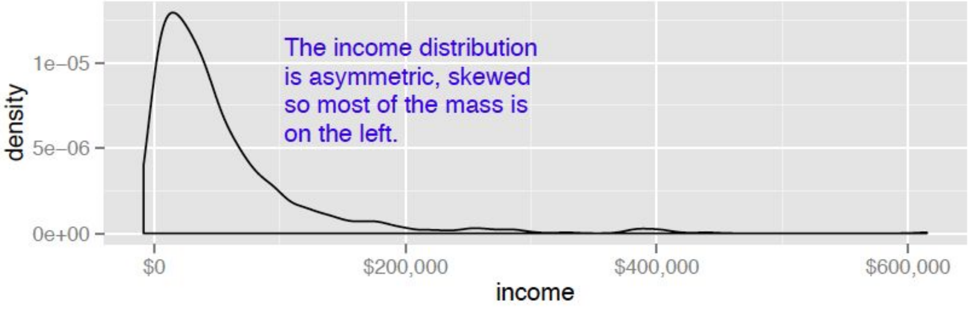
- “ln” denotes a **natural log**. The base is *e* (approx. 2.72).

$\ln 100$ is the same as $\log_e 100 \;\; \approx 4.61$

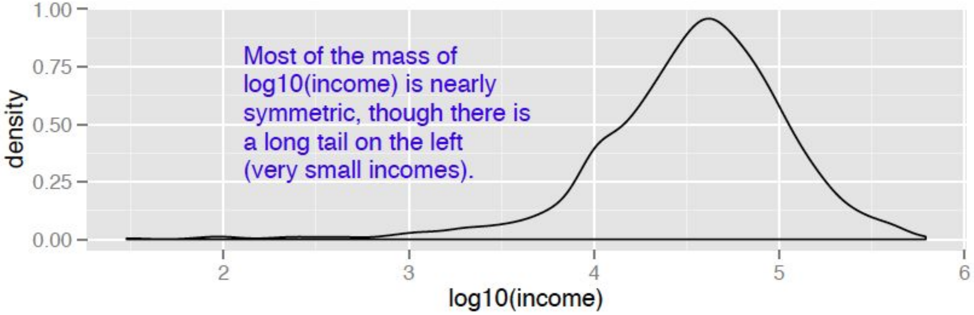
One Statistics Application of Logs (among many!):

- Using log transformations can help to make a skewed distribution look more "Normal"

  Instead of plotting a histogram of each person's income, we can take the log of each person's income and plot that.



The income distribution is asymmetric, skewed so most of the mass is on the left.



Most of the mass of log10(income) is nearly symmetric, though there is a long tail on the left (very small incomes).

Exponent Properties

1. $a^x a^y = a^{x+y}$

2. $a^{-x} = \frac{1}{a^x}$

3. $a^{xy} = (a^x)^y$

4. $a^0 = 1$

Logarithm Properties

1. $\log_a(xy) = \log_a(x) + \log_a(y)$

2. $\log_a(x^y) = y\log_a(x)$

3. $\log_a(x)$ when $x \leq 0$ is undefined

Practice with exponents and logarithms

Presenting Data

Representing data visually using Stata:

1. Download the updated dataset (“BootcampDay2.dta”) from the Box folder and save it in your working directory from yesterday.

2. Open your do file from yesterday. Change the code so that your do file opens “BootcampDay2.dta” rather than “Workshop.dta”.

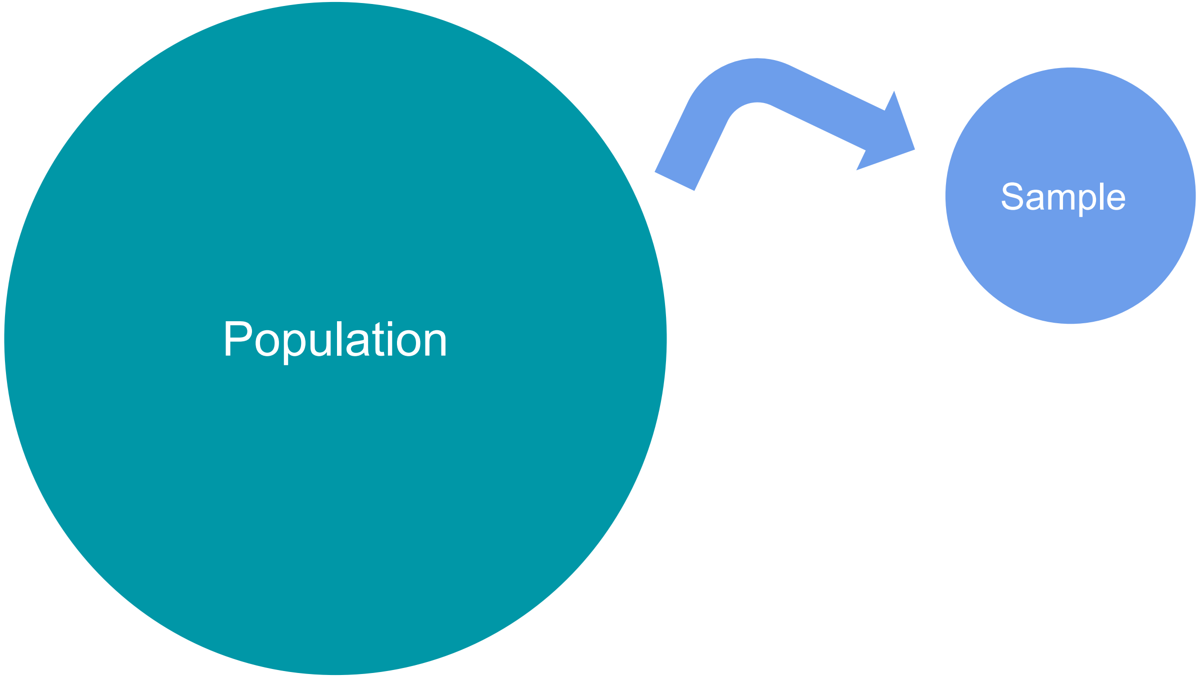
3. Complete the self-directed handout from yesterday (through activity 5).

4. Try to recreate the graph(s) you drew by hand using Stata. ***If you make a graph you are really proud of, email it to us! (rgleit@stanford.edu)***

lunch

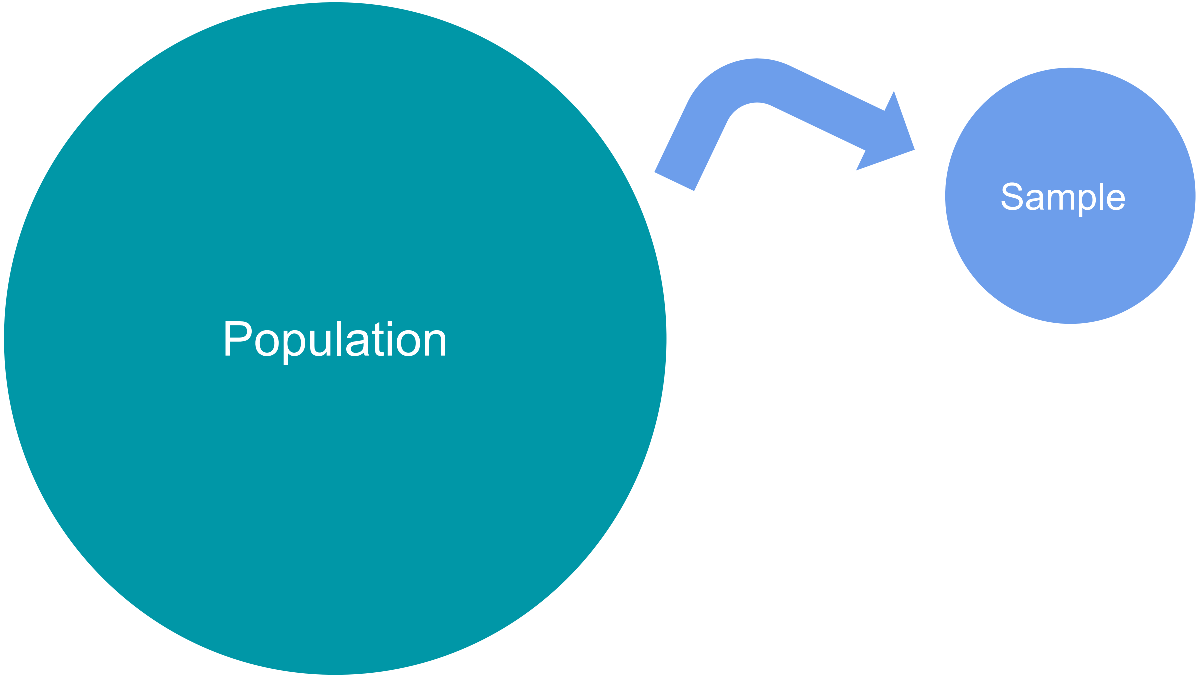
Review

What is the difference between **descriptive statistics** and **inferential statistics**?



Review

What is meant by **signal** vs. **noise**?



Elgar

Practice: Distributions — Part A Questions

- What do you notice about the three plots you made?
- What is the difference between what you plotted in 1, 2, and 3?

Practice: Distributions — Part B Questions

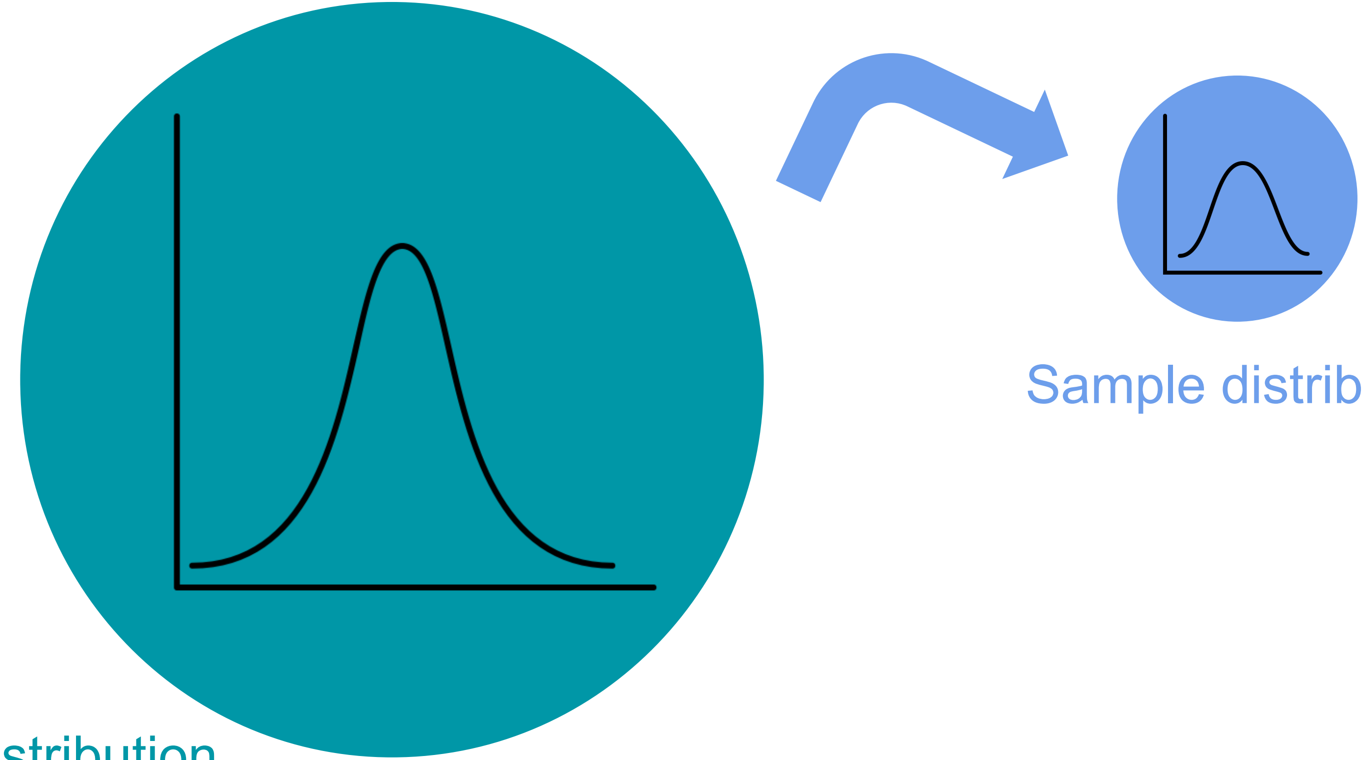
- What was different about the plots in part B compared to part A?
- What was similar about the plots in part B compared to part A?

Distributions

- What is a distribution?

- 3 types:
  - Population distribution
  - Sample distribution
  - Sampling distribution

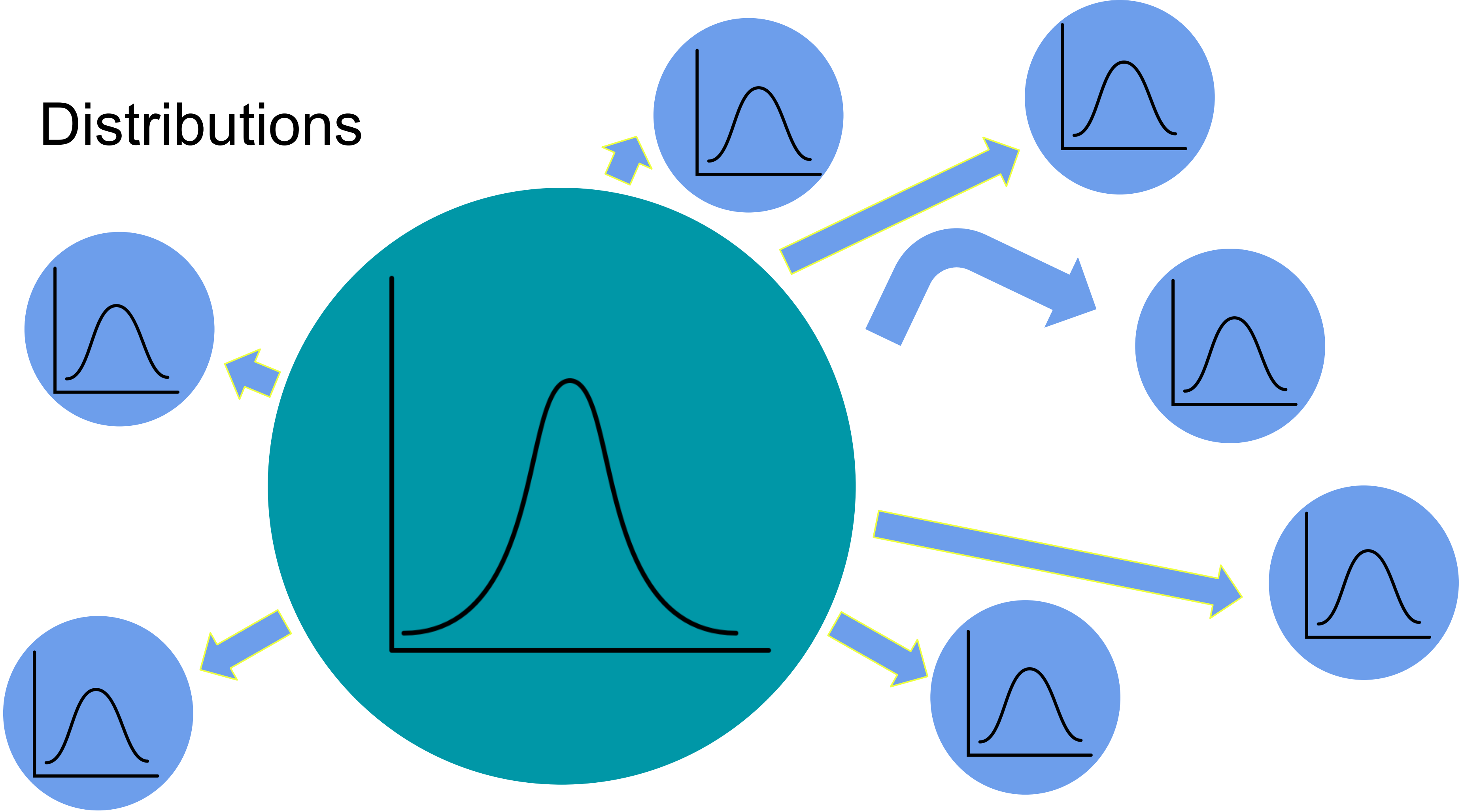
Distributions



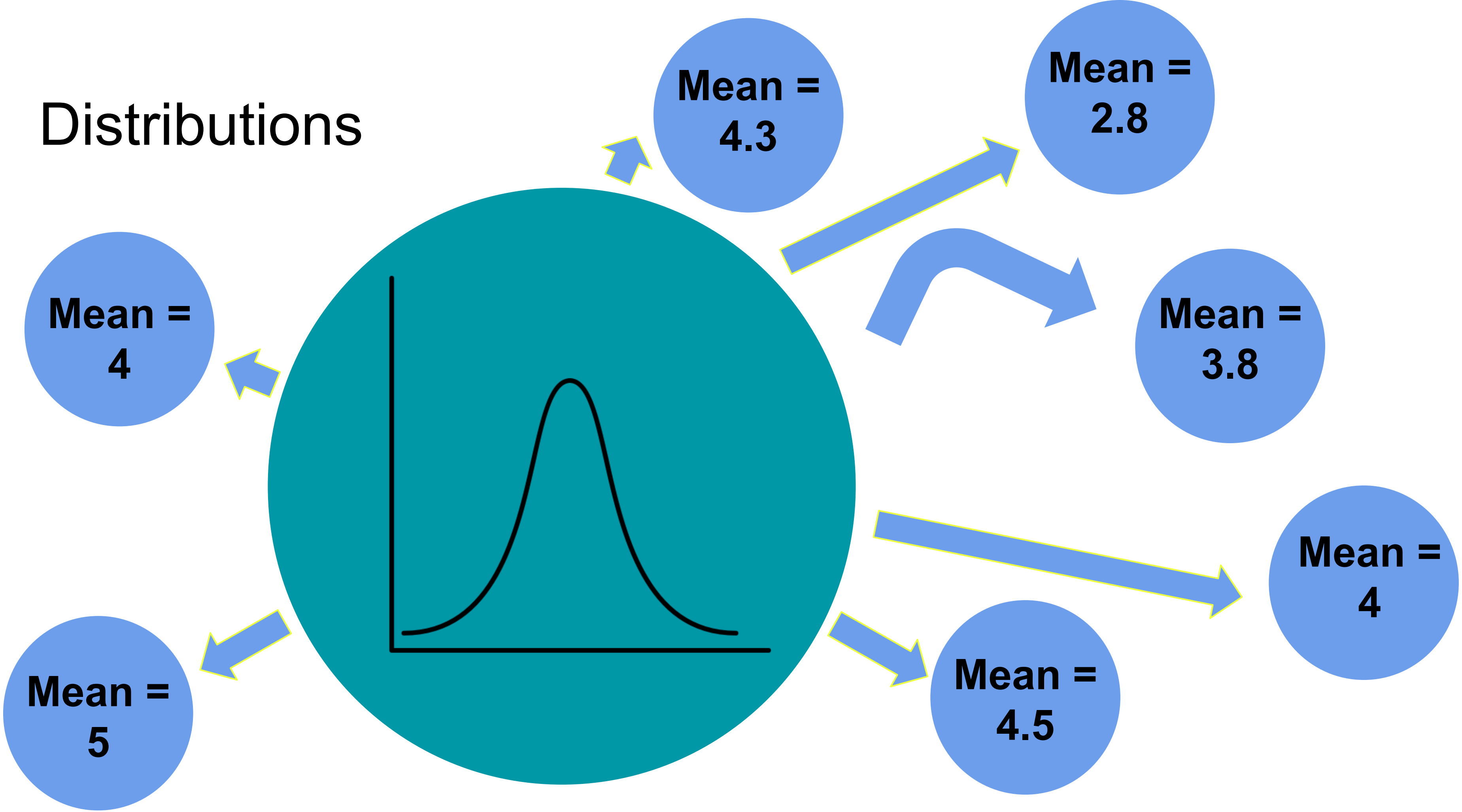
Sample distribution

Population distribution

Distributions



Distributions



Mean = 4

Mean = 4.3

Mean = 2.8

Mean = 3.8

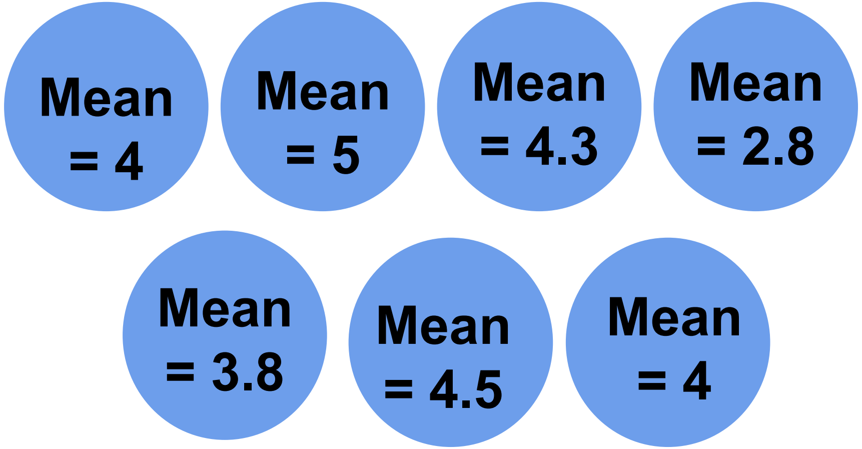
Mean = 4

Mean = 4.5

Mean = 5

Sampling Distribution

The distribution of means



Mean = 4

Mean = 5

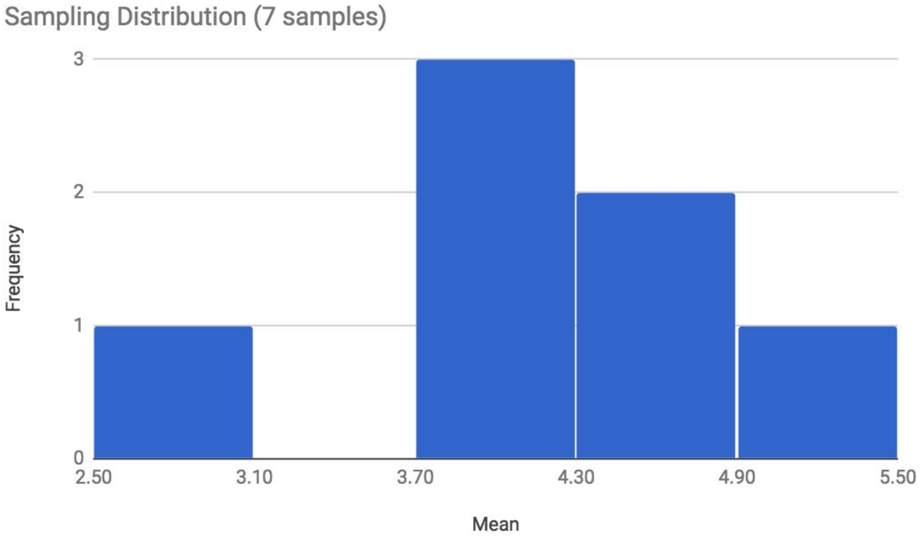
Mean = 4.3

Mean = 2.8

Mean = 3.8

Mean = 4.5

Mean = 4

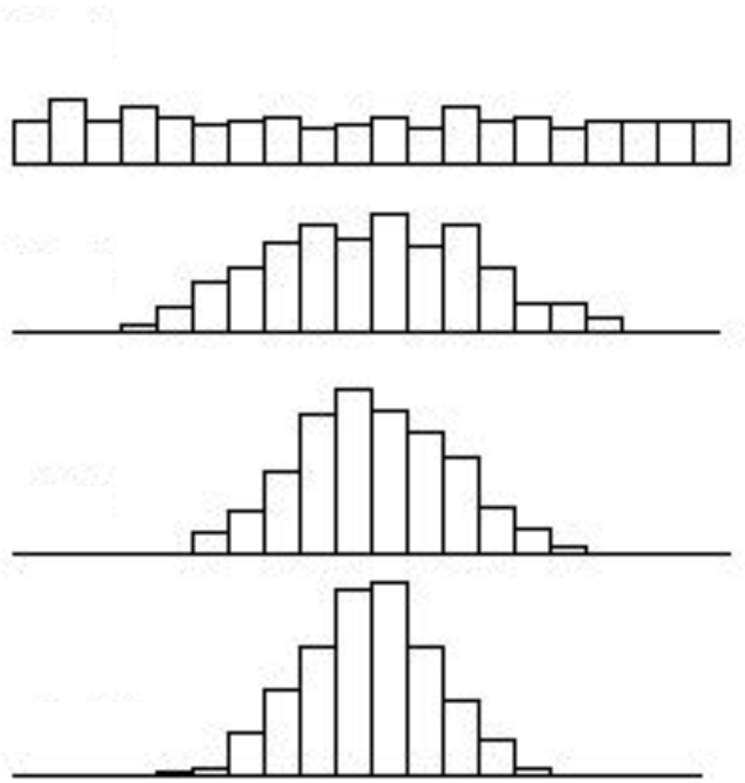


Sampling Distribution

- What happens to the sampling distribution as the number of repeated samples increases?

- What happens to the sampling distribution as the number of observations in each sample increases?

Sampling Distribution



https://istats.shinyapps.io/sampdist_cont/

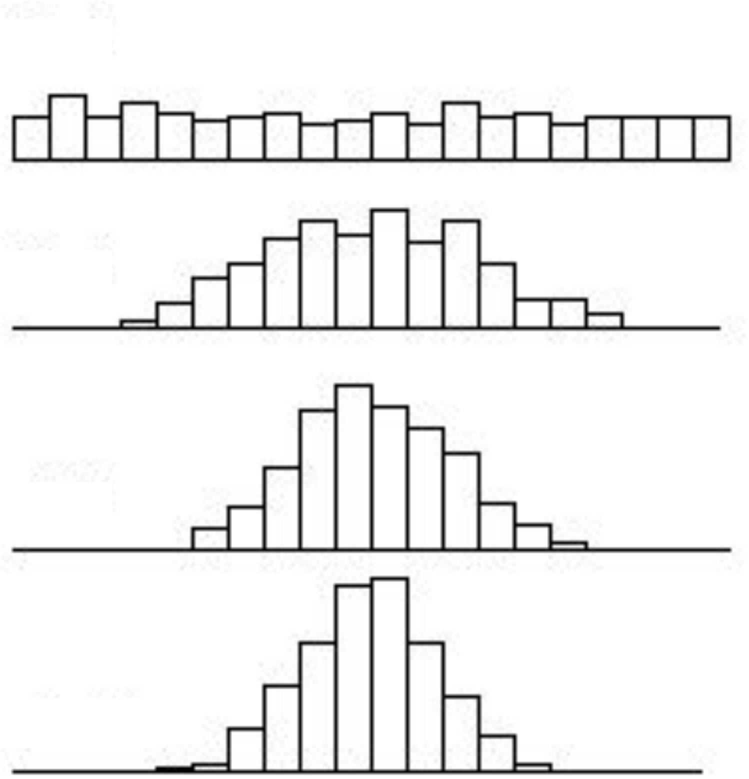
Sampling Distribution

Sample size:
- Tangible
- You control this
- Bigger is better (>30)

100

100

100



100

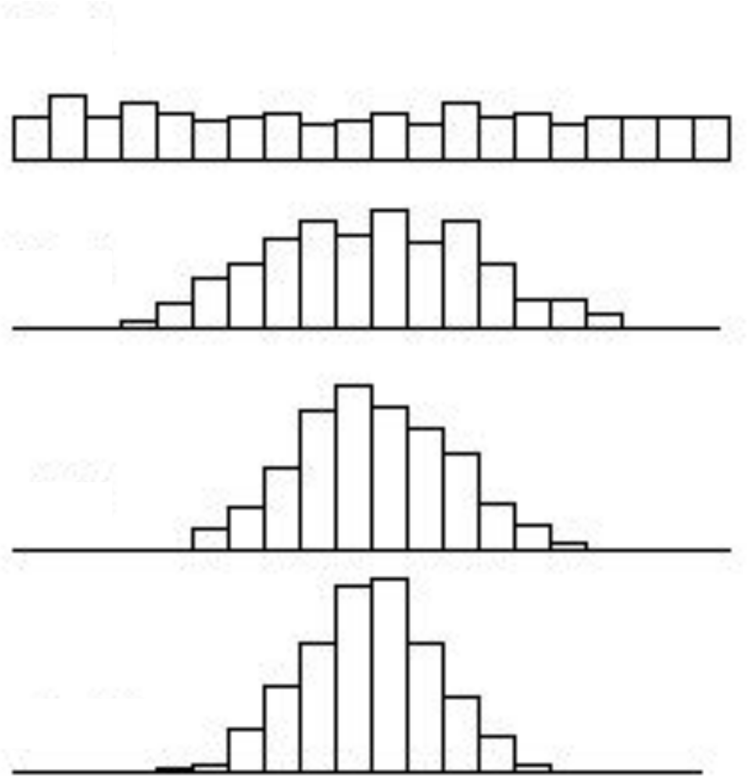
https://istats.shinyapps.io/sampdist_cont/

Sampling Distribution

Sample size:
- Tangible
- You control this
- Bigger is better (>30)

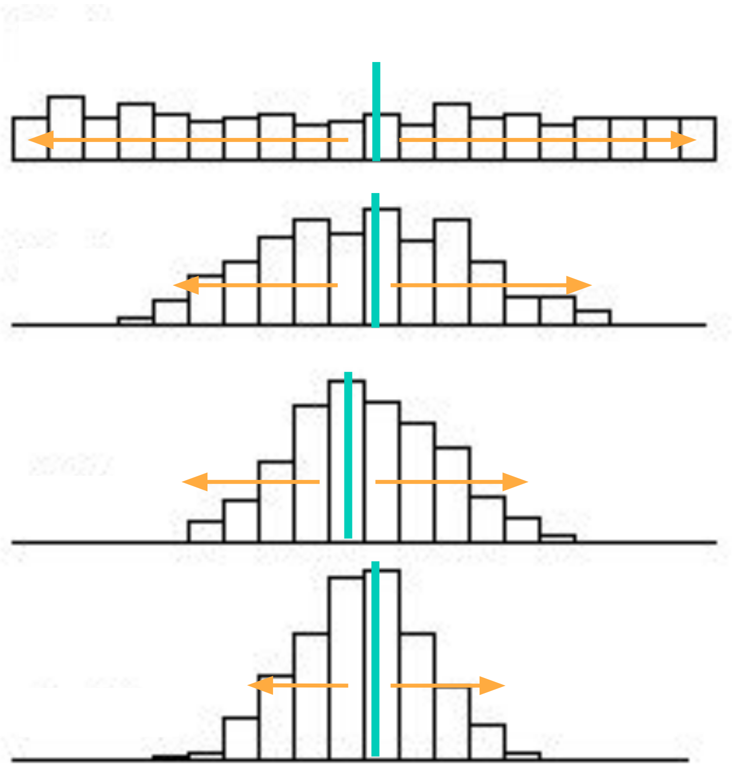
# of repeated samples:
- Imagined
- As it gets infinitely large, the distribution has nice statistical properties

20	100
40	100
80	100
160	100



https://istats.shinyapps.io/sampdist_cont/

Signal vs. Noise in the Sampling Distribution



Signal vs. Noise in the Sampling Distribution

Mean → ***estimate***

Standard deviation → ***standard error***