Managing {Probability, Projects}

September 18, 2018
Statistics Bootcamp

Agenda

- Set notation and probability
- English 🔄 Probability notation
- The Sampling: Part II
- Hypotheses and Inferences
- Lunch
- Organizing your Stata/mind

Warm up with hot takes

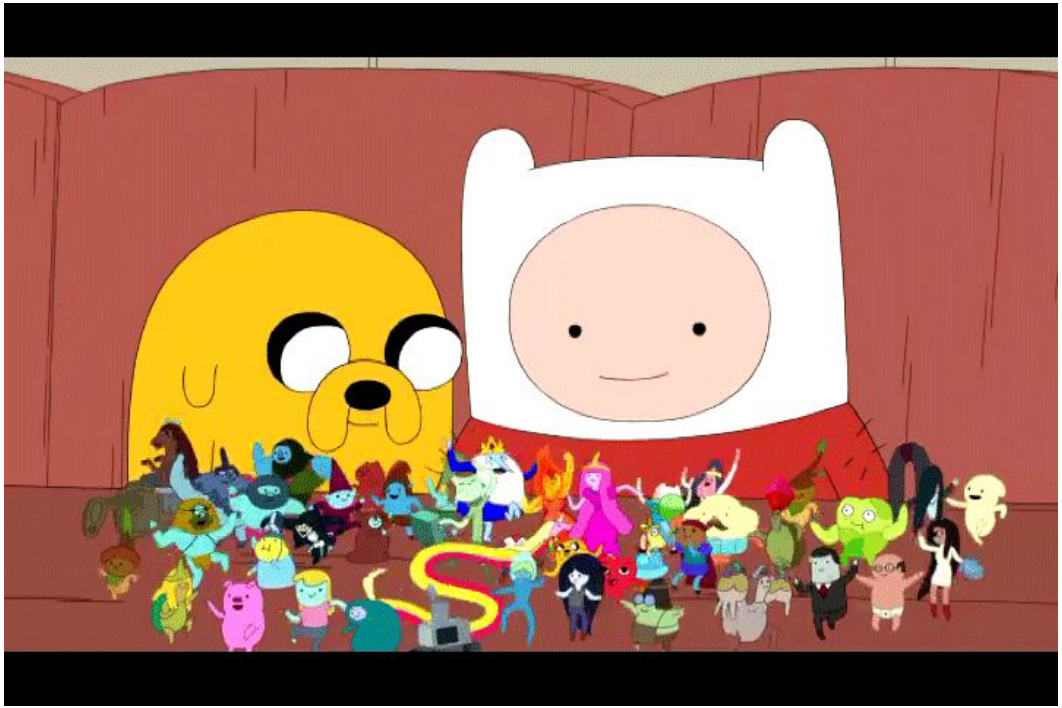
- Write a hot take
  - You don’t have to believe it
  - It can be cold
- Describe evidence that would signal your take is correct
- Describe noise that would make people less likely to accept your take

Picture telephone

Goals for Probability and Distributions

- Practical:
  - ...read and use probability notation to describe a situation with bounded uncertainty.
- Conceptual:
  - ...explain some ways that probabilistic model assumptions can undermine our models of society.

Statistical models, probability, and assumptions



“It’s like a finger pointing a way to the moon….

Don’t concentrate on the finger,

Or you will miss all the heavenly glory.”

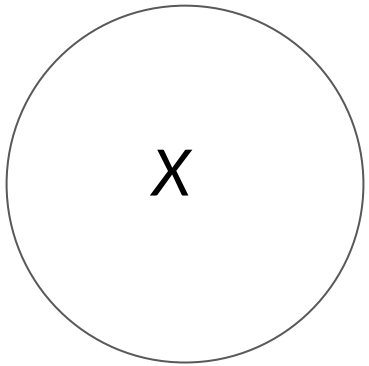
Bruce Lee, Enter the Dragon, 1973

Probability and set notation

The set of all events: {}

Any given day can be rainy, sunny, or surreal.

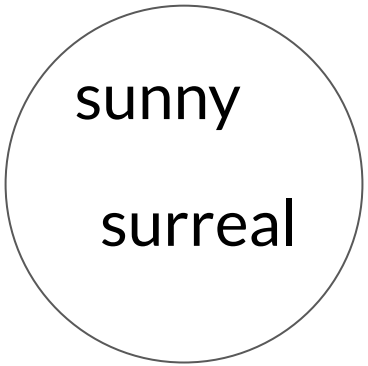
Possible weather events = {rainy, sunny, surreal} = *X* or some other capital letter



The set of all X such that: {X: some condition}

Imagine the set of all weather events such that you could get a tan.

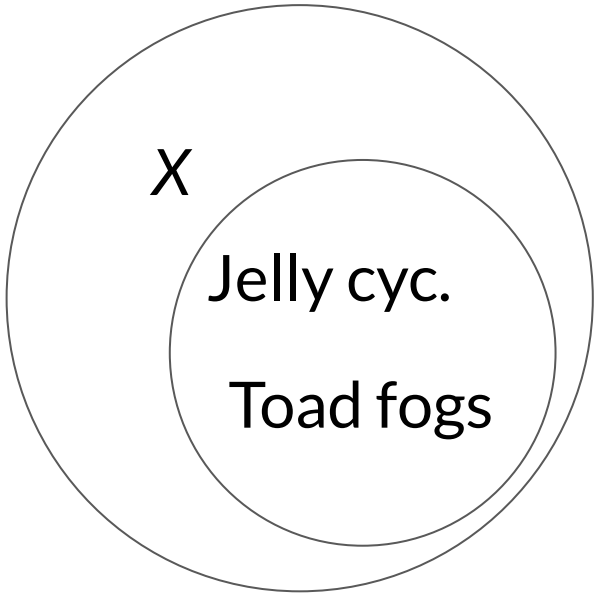
{Weather: you could get a tan} = {sunny, surreal}



Is an element of: ∈

Jelly cyclones and toad fogs are elements of surreal weather.

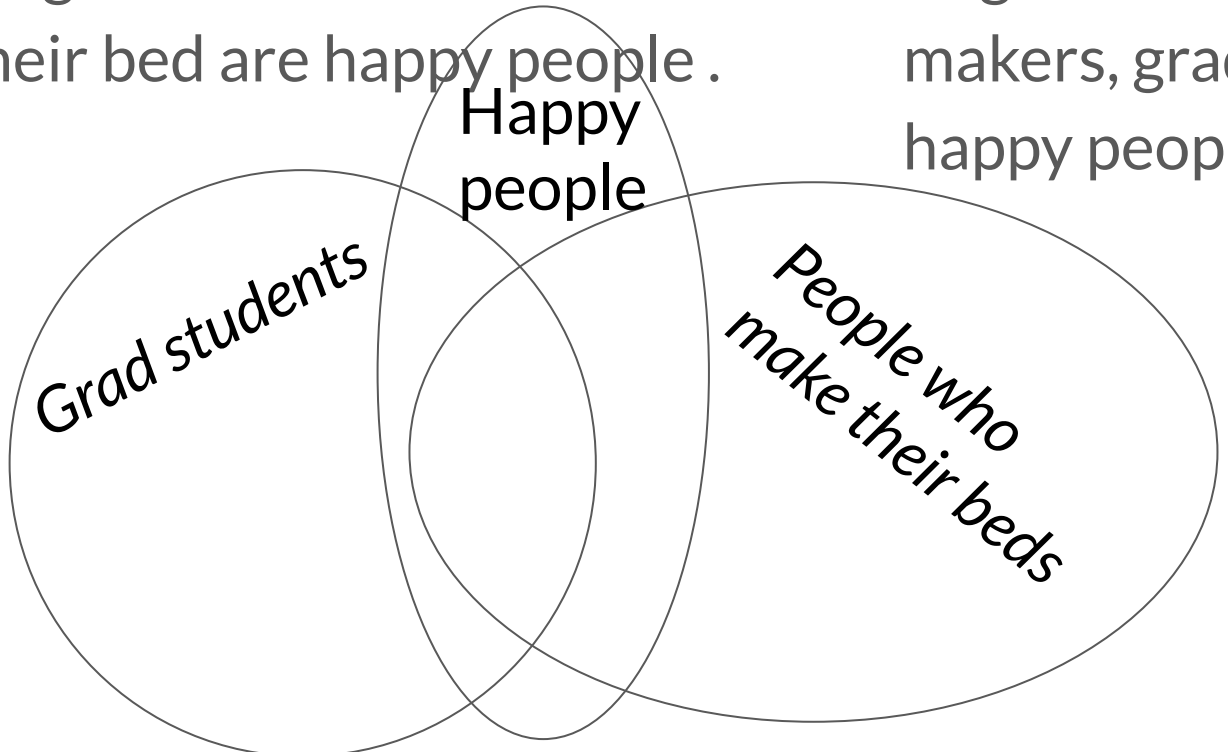
{jelly cyclones, toad fogs} ∈ Surreal weather ∈ Weather



For all: ∀

All grad students who make their bed are happy people .

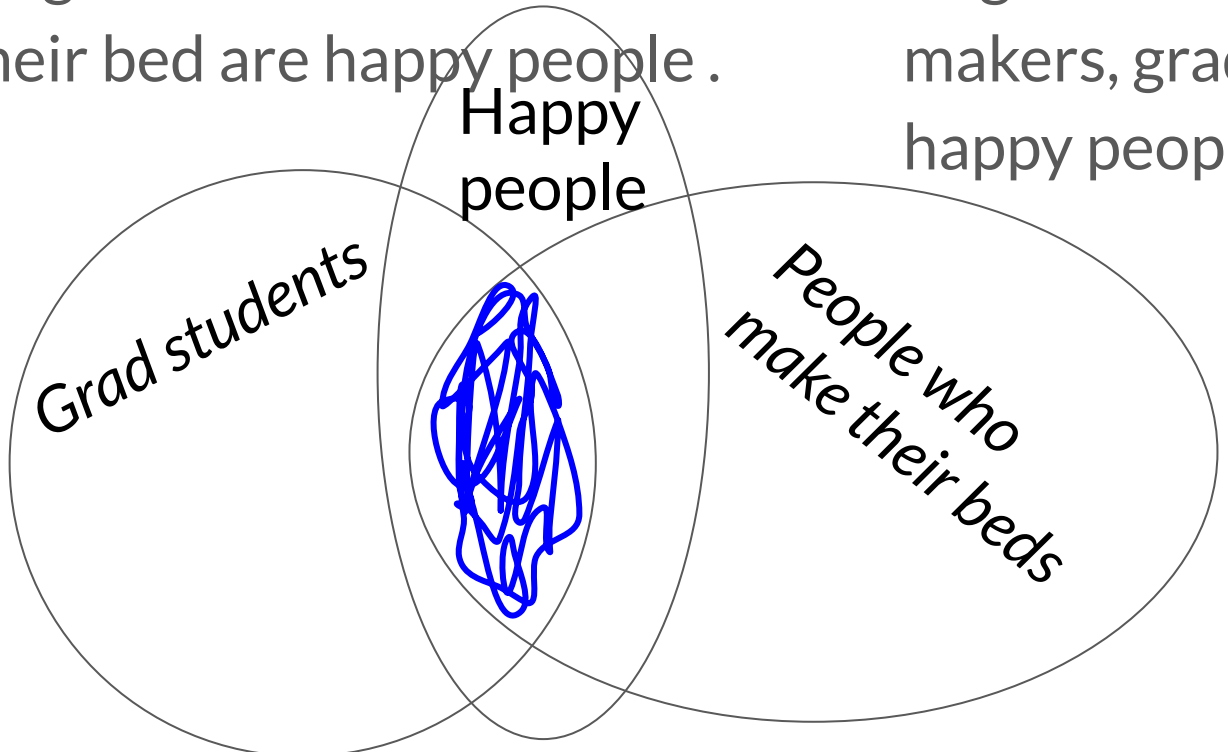
∀ grad students ∈ bed makers, grad students ∈ happy people



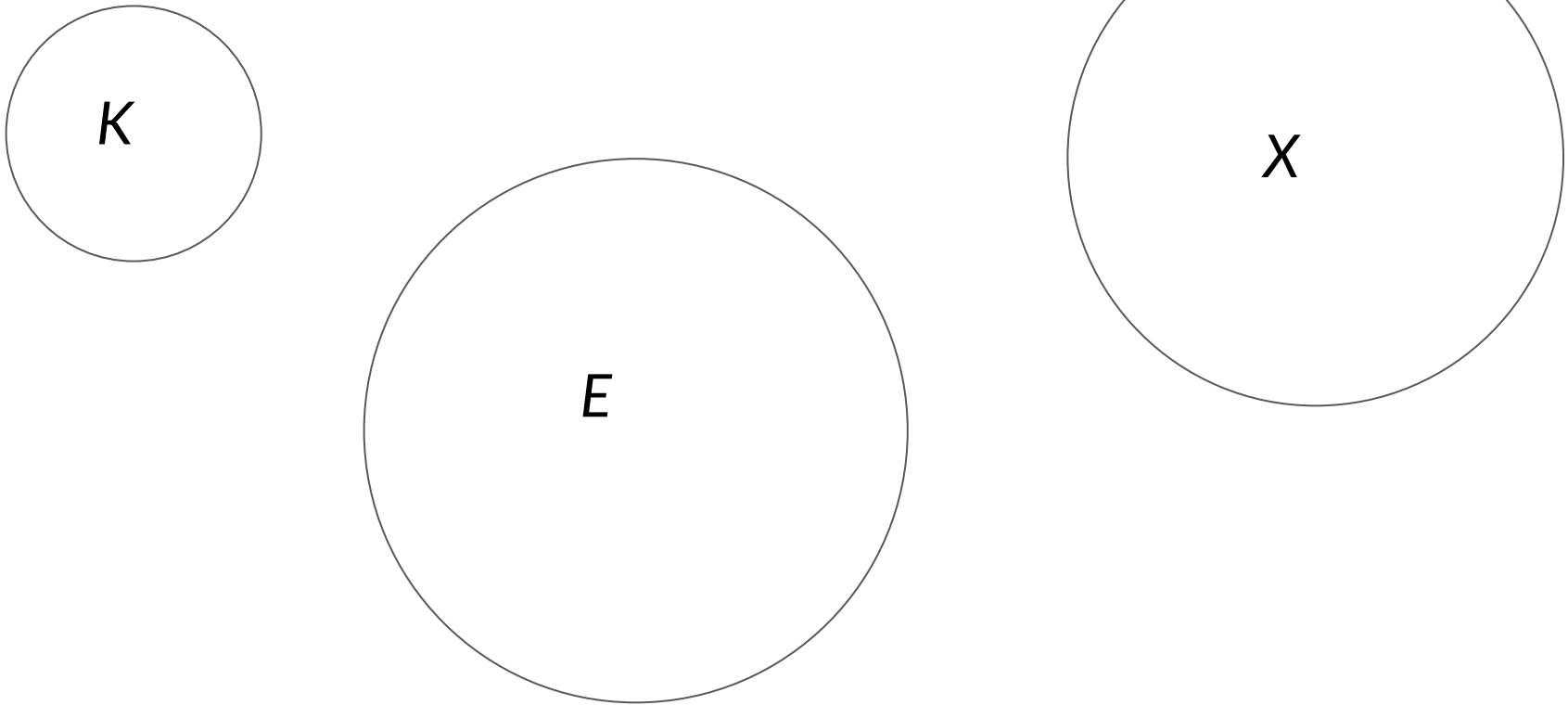
For all: ∀

All grad students who make their bed are happy people .

∀ grad students ∈ bed makers, grad students ∈ happy people



Exercise



Write a set notation phrase to the neighbor on your left
Translate!

Flip index card, write a plain-english phrase to right neighbor
Translate!

The probability that an event occurs: Pr(x)

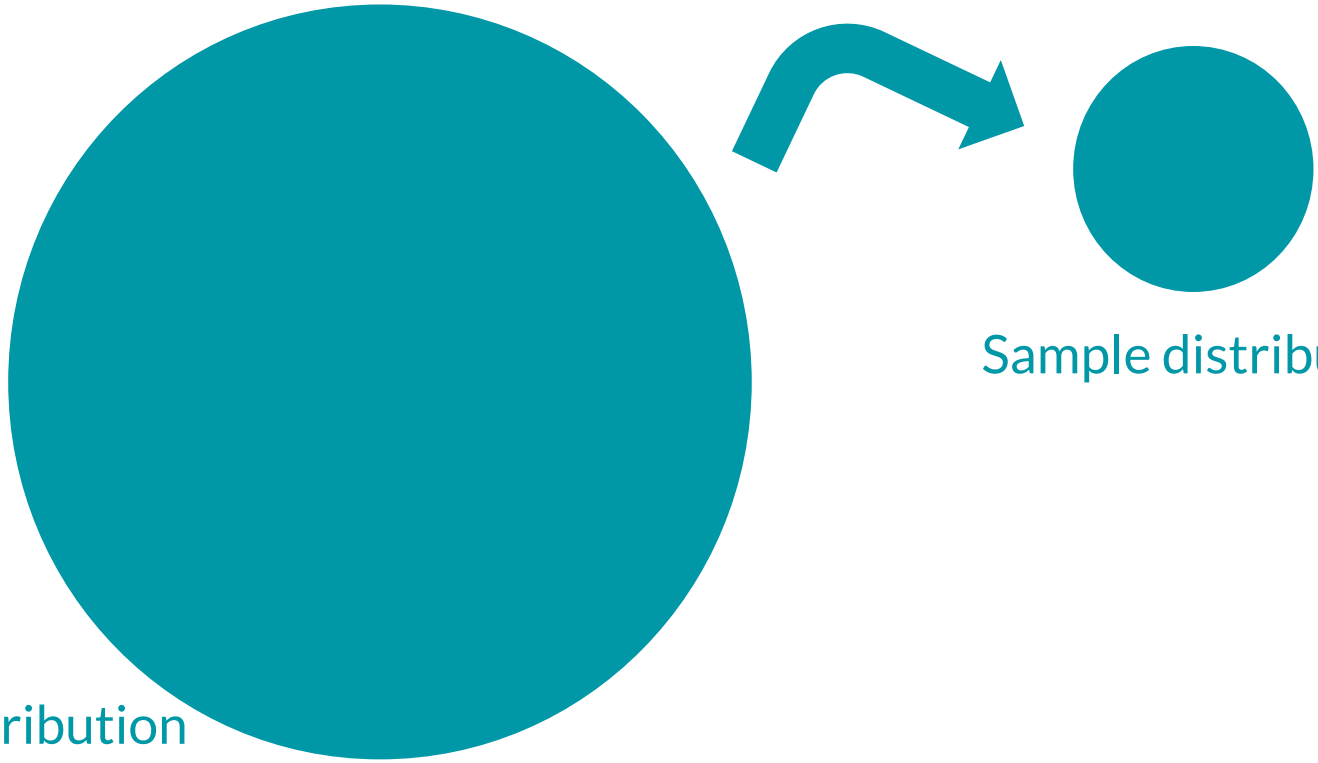
There is a 70% chance of rain.

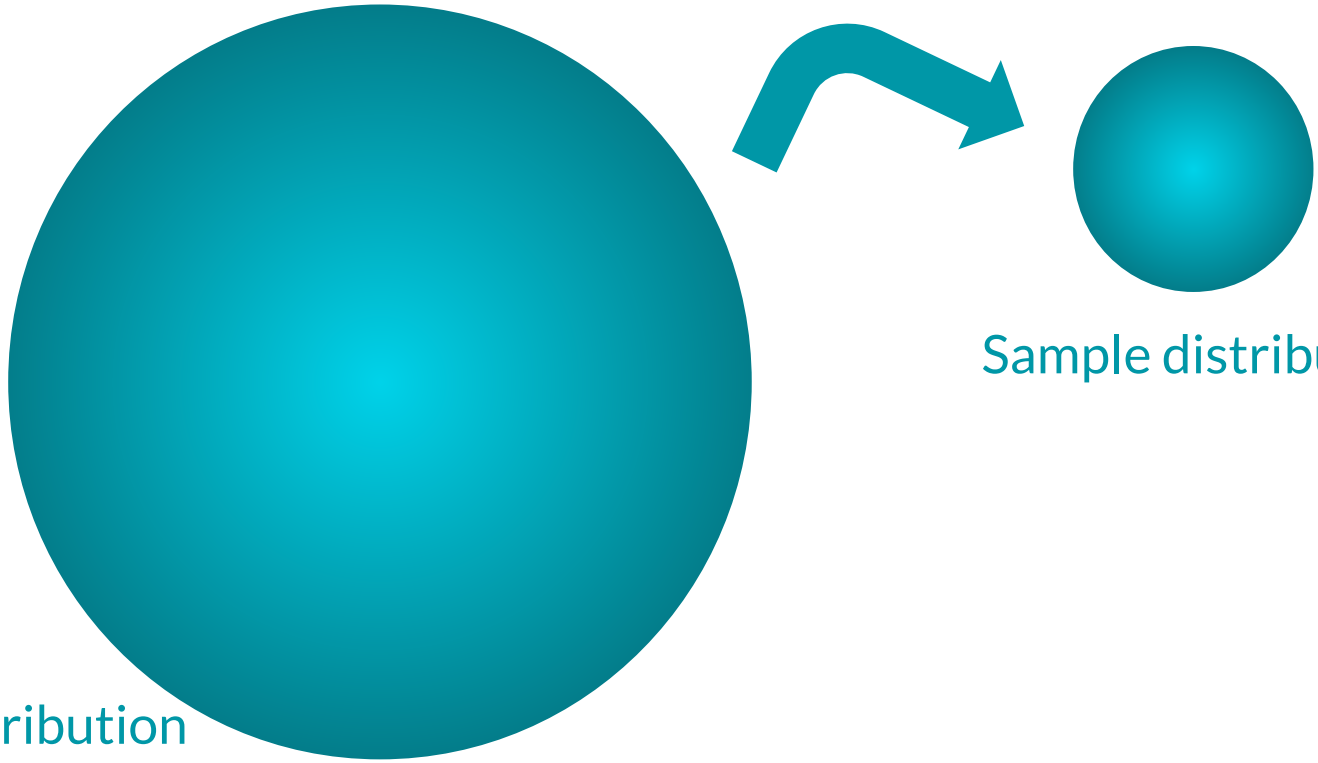
Pr(rain) = 0.7

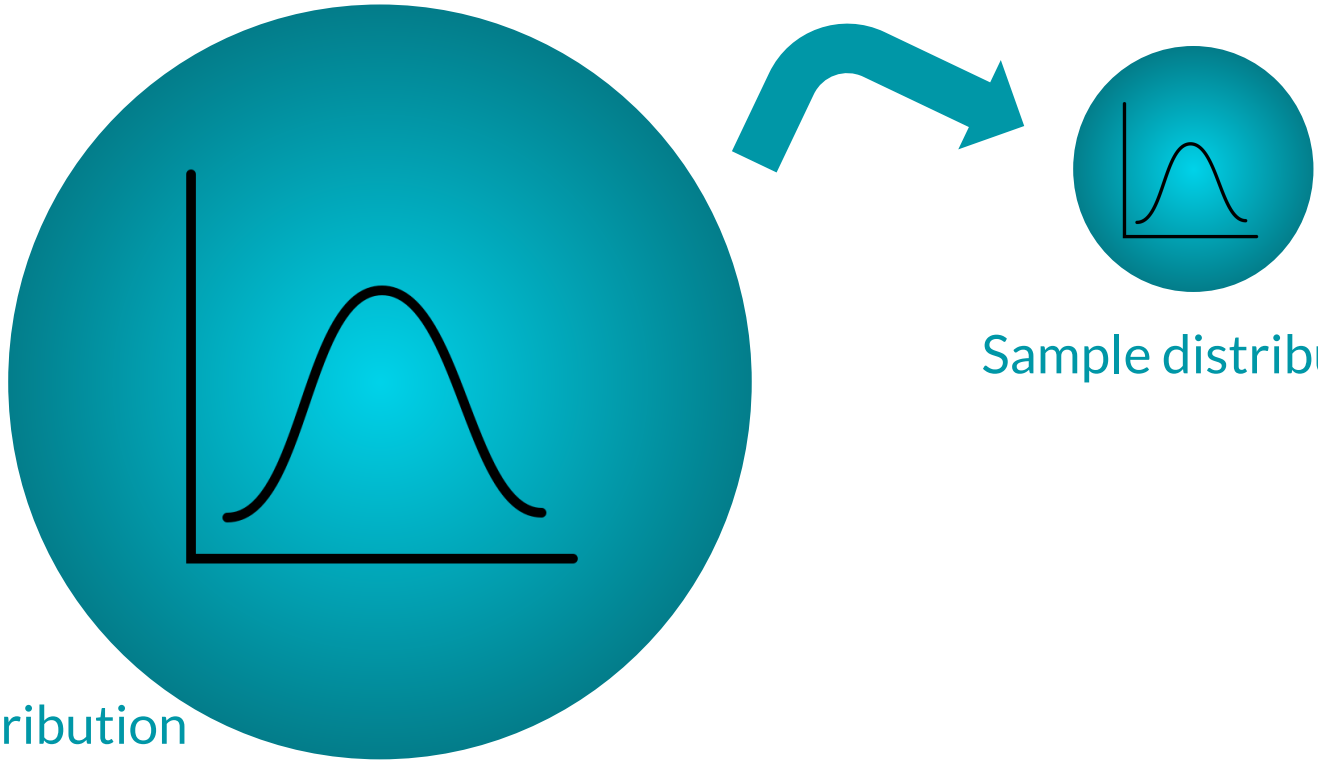
Pr(no rain) = 0.3

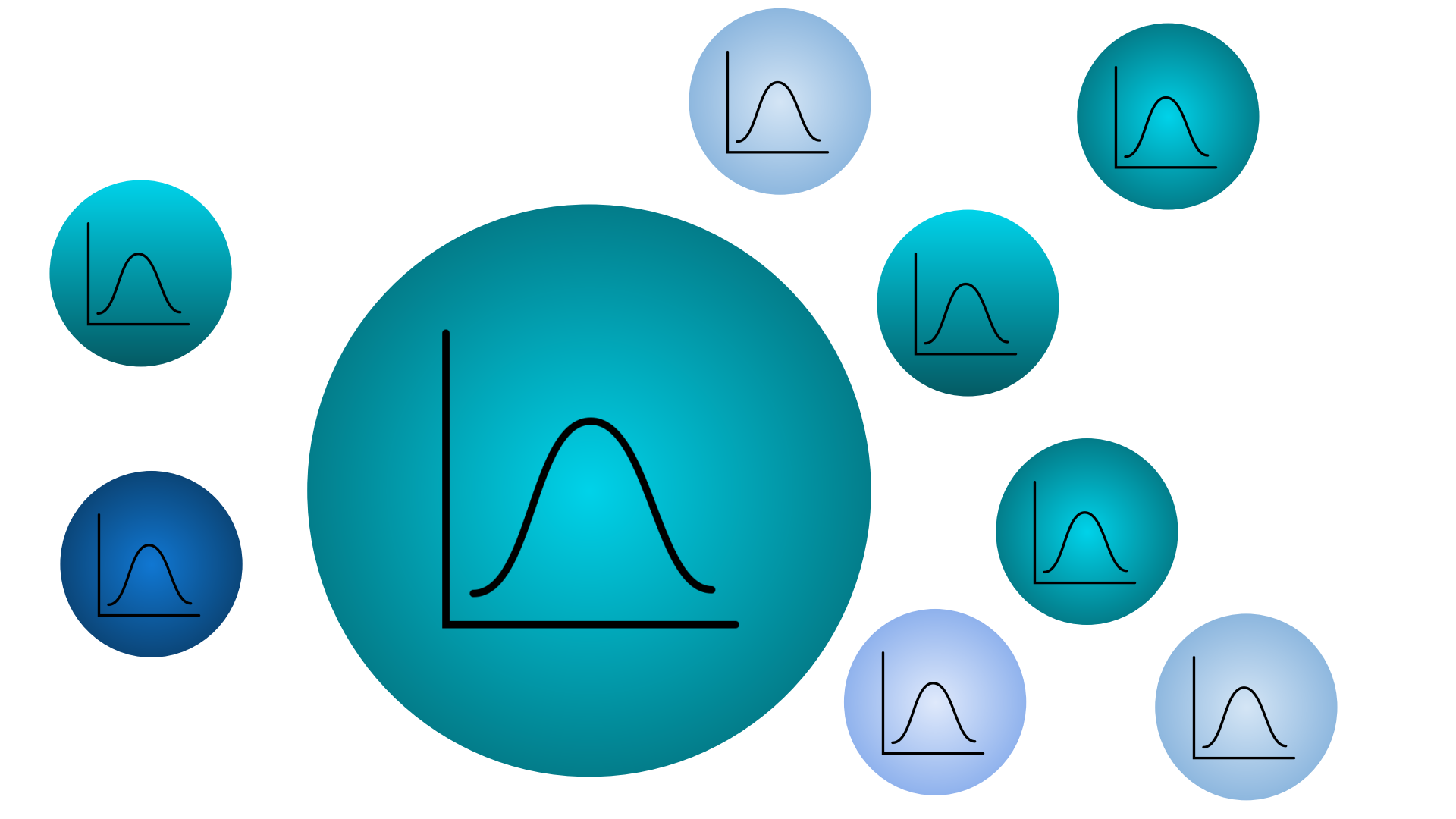
Pr(weather no matter how surreal) = 1

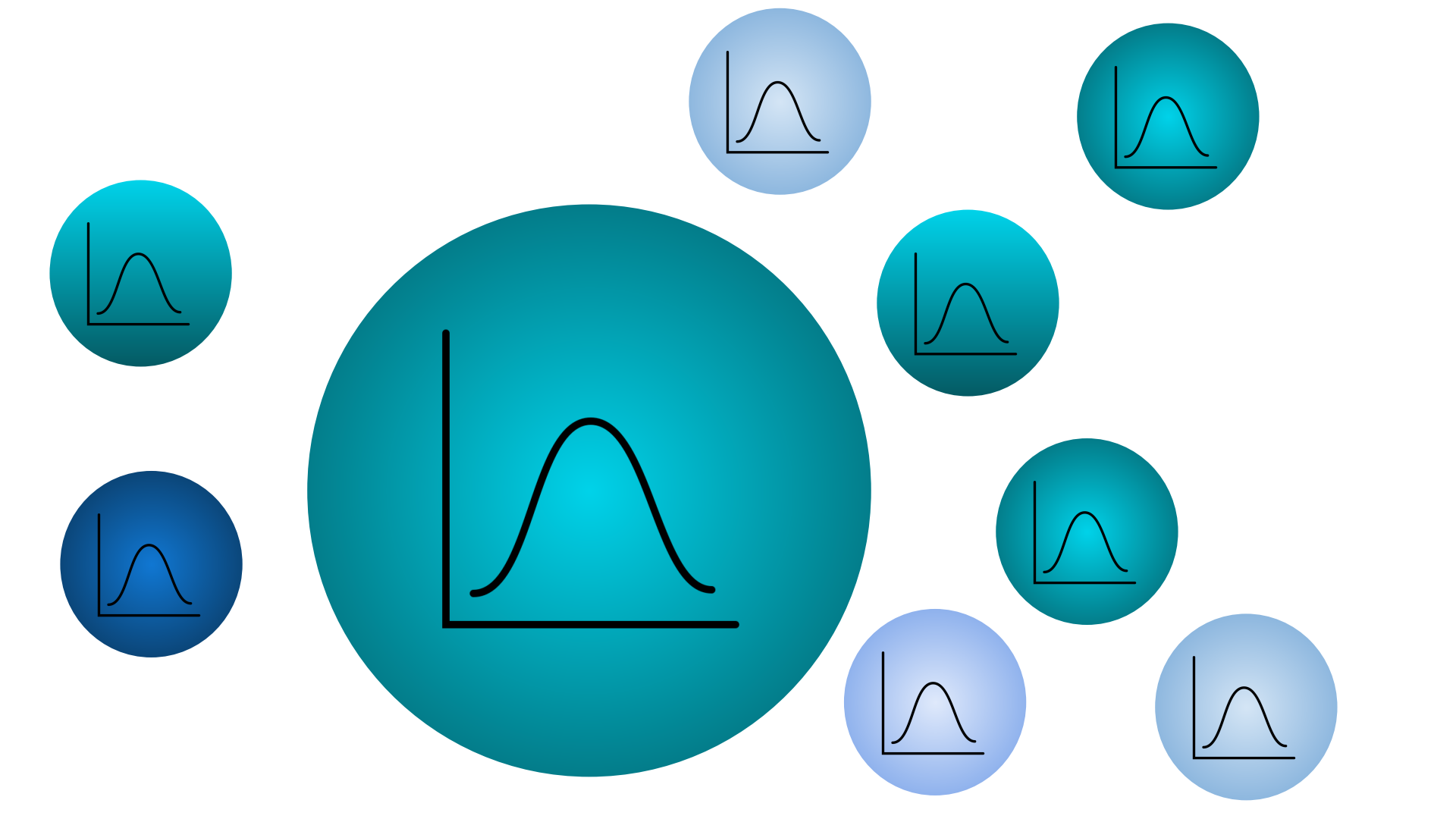
REVIEW: Samples, distributions, and the Sampling

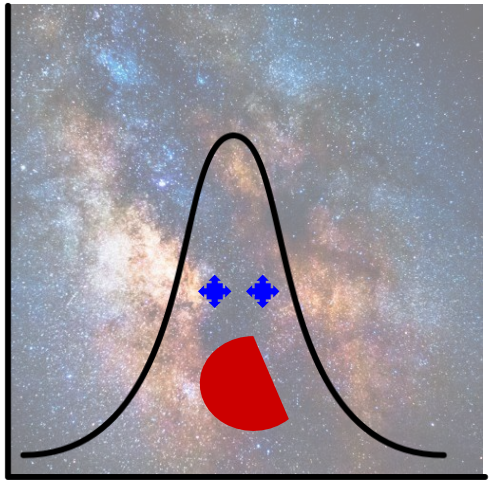

Sample distribution
Population distribution


Sample distribution
Population distribution

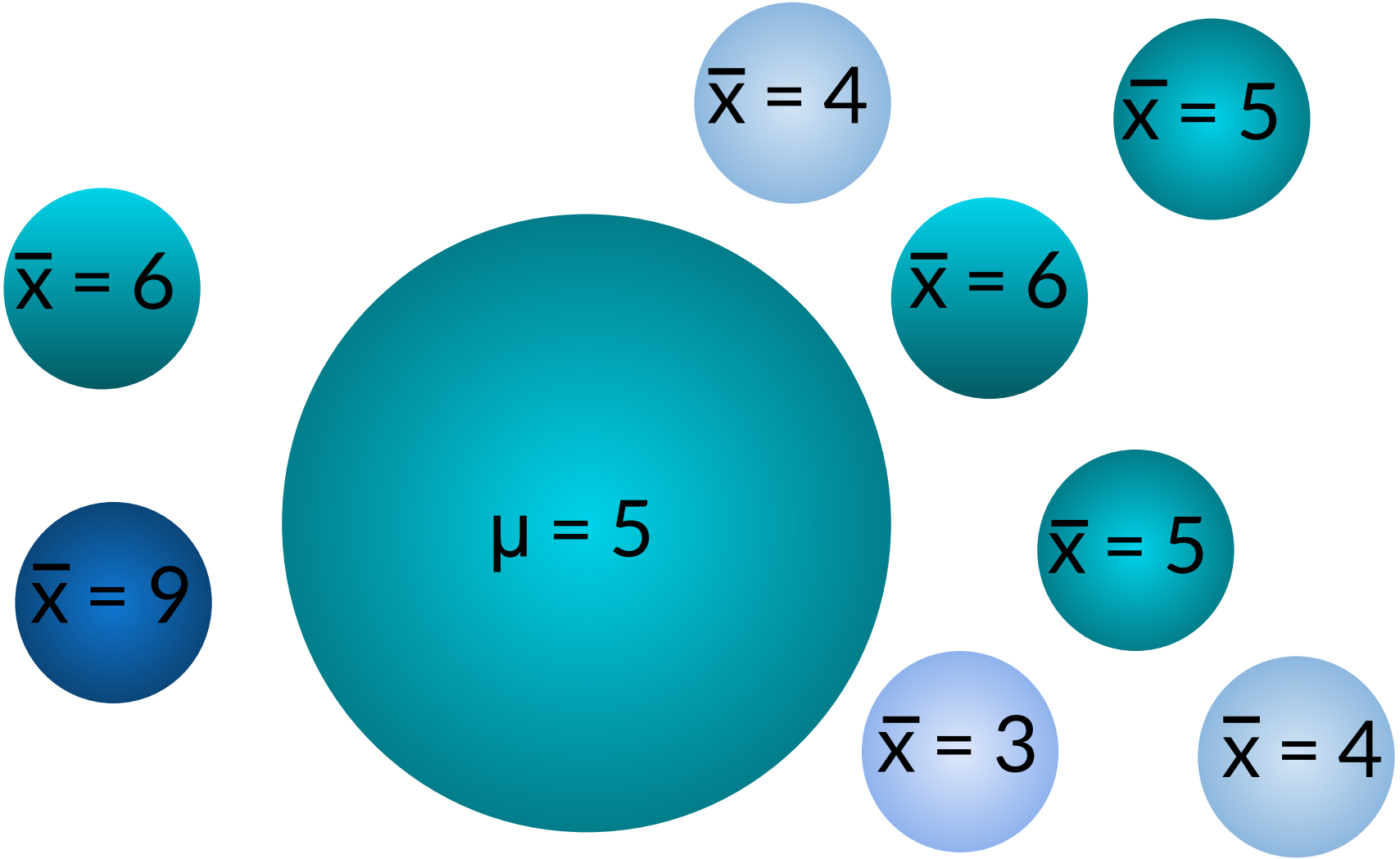

Sample distribution
Population distribution

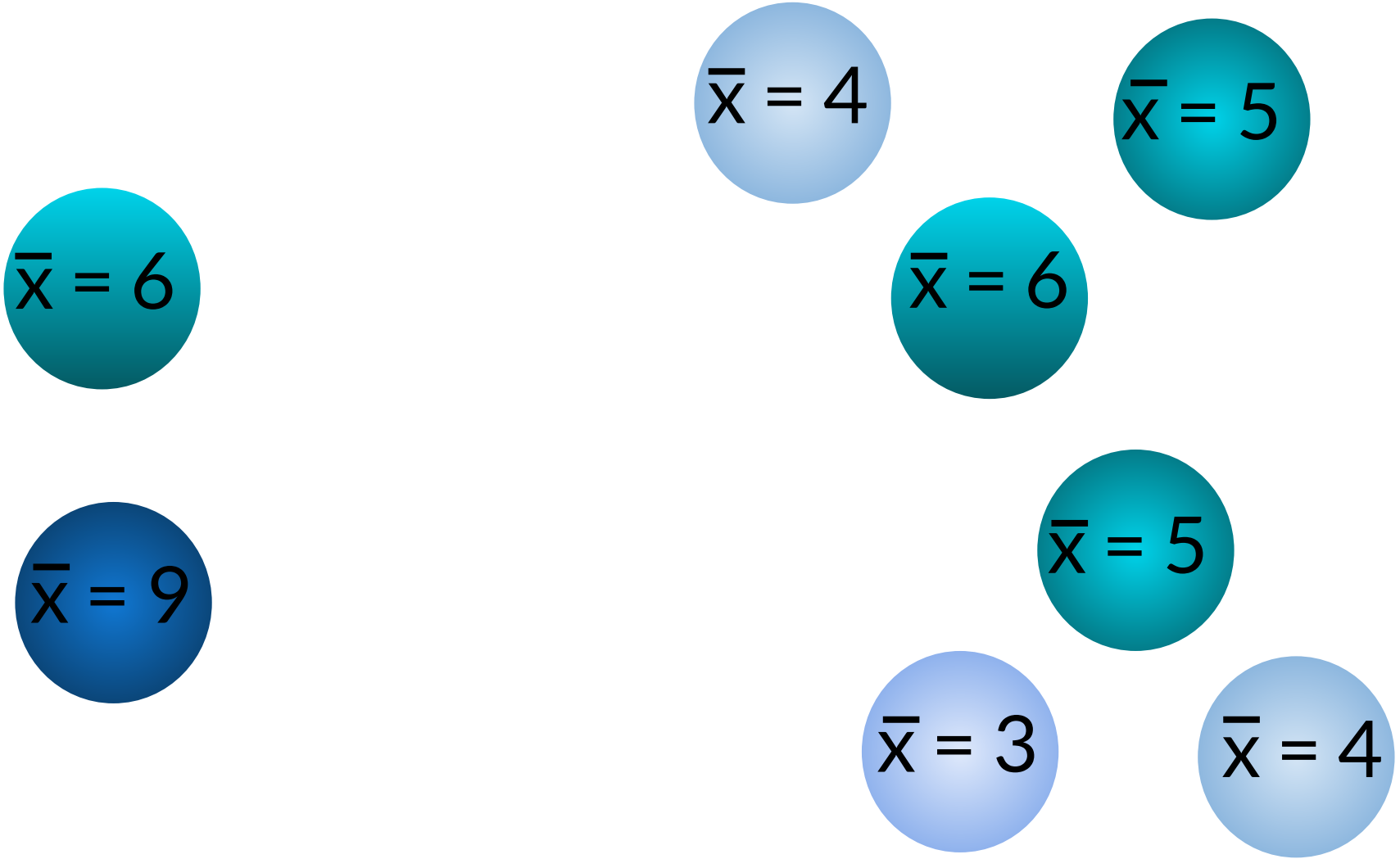


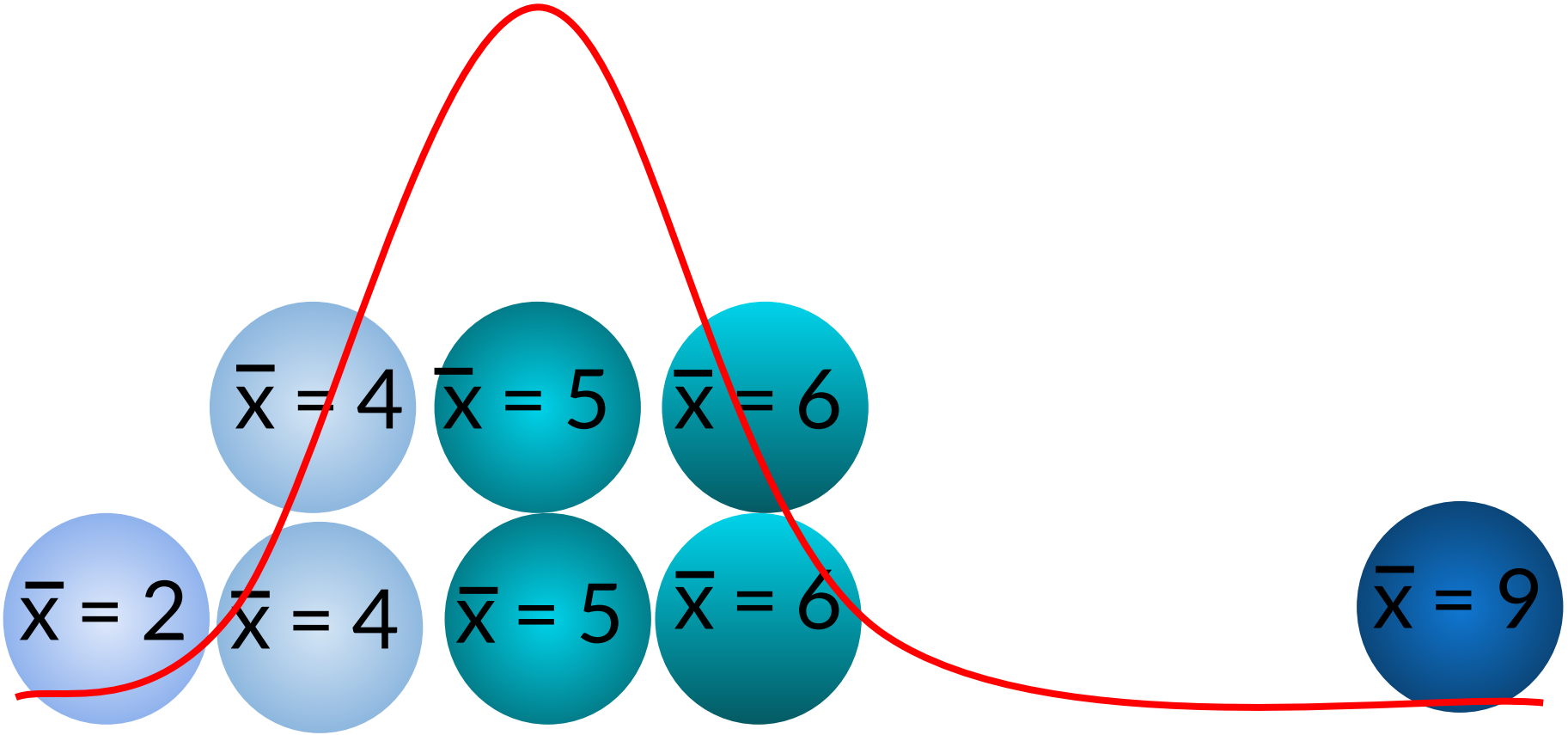


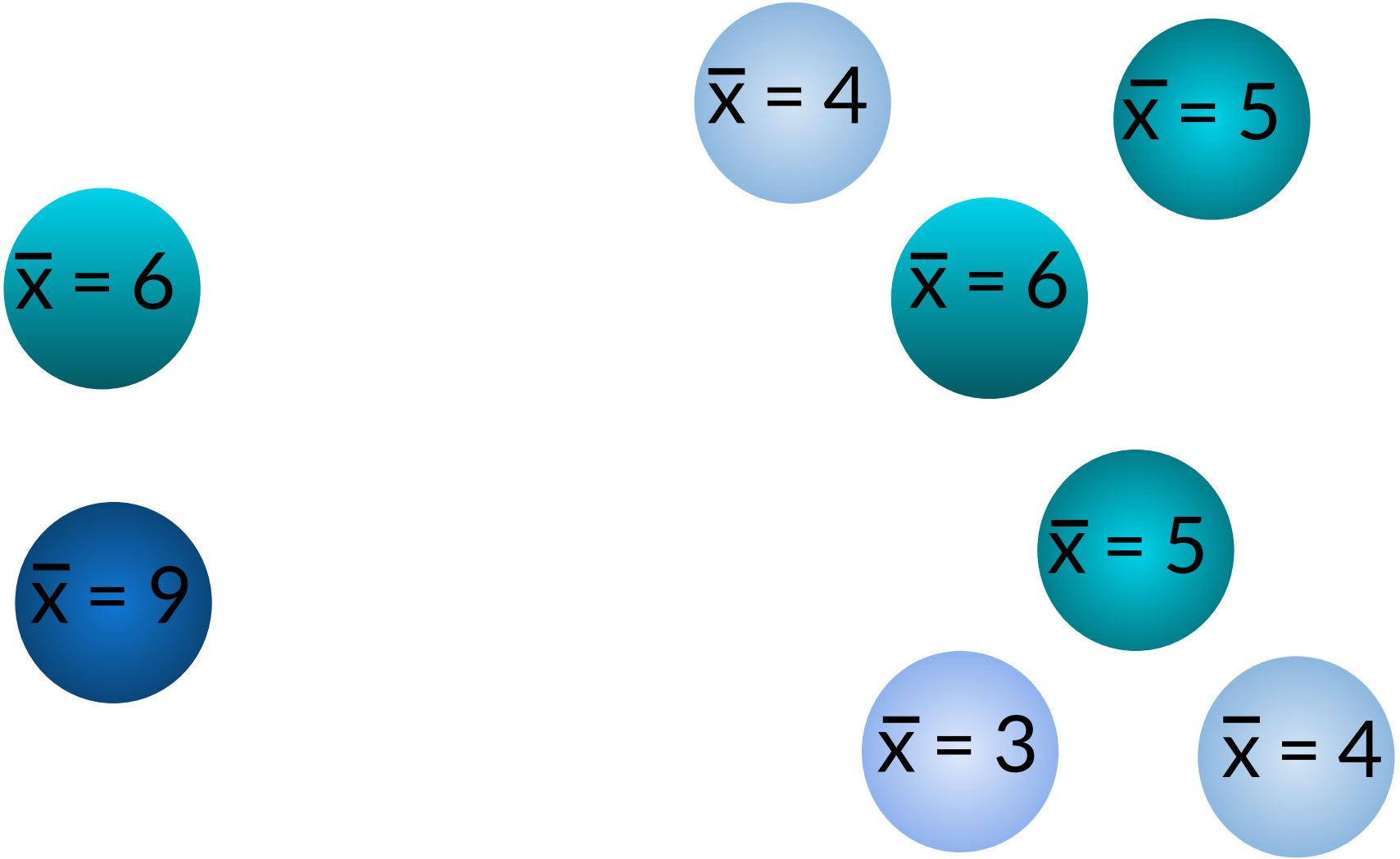


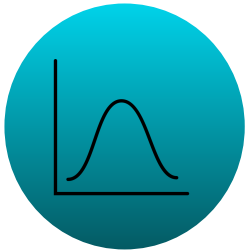
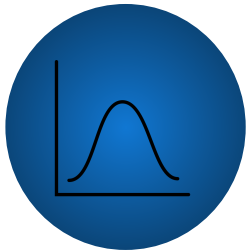
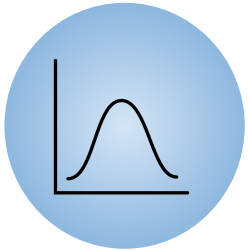
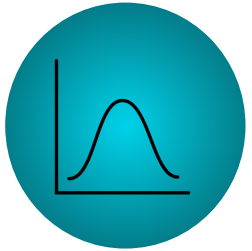
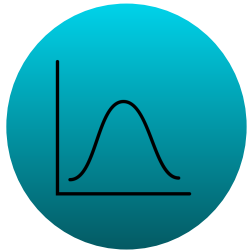
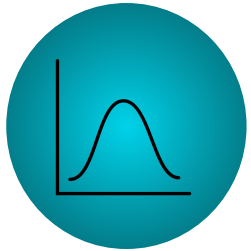
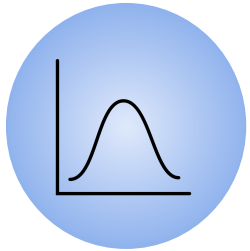
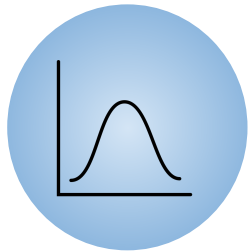
Sampling distribution


$\bar{x} = 4$
$\bar{x} = 5$
$\bar{x} = 6$
$\bar{x} = 6$
$\mu = 5$
$\bar{x} = 9$
$\bar{x} = 5$
$\bar{x} = 3$
$\bar{x} = 4$


$\bar{x} = 4$
$\bar{x} = 5$
$\bar{x} = 6$
$\bar{x} = 6$
$\bar{x} = 9$
$\bar{x} = 5$
$\bar{x} = 3$
$\bar{x} = 4$


$\bar{x} = 4$
$\bar{x} = 5$
$\bar{x} = 6$
$\bar{x} = 2$
$\bar{x} = 4$
$\bar{x} = 5$
$\bar{x} = 6$
$\bar{x} = 9$


$\bar{x} = 4$
$\bar{x} = 5$
$\bar{x} = 6$
$\bar{x} = 6$
$\bar{x} = 9$
$\bar{x} = 5$
$\bar{x} = 3$
$\bar{x} = 4$

http://onlinestatbook.com/stat_sim/sampling_dist/

Or search: `onlinestatbook sampling`

5m Questions:

1. How does the sampling distribution compare to the population distribution (normal/uniform/skewed)?
2. What happens when we change the sample size used to generate the sampling distribution?
3. What custom population distribution produces a non-normal sampling distribution?

Now we want to know if dogs are cuter than cats.

What could we do?

Now we want to know if dogs are cuter than cats.

We could:

1. Get two samples: one of dogs and one of cats.
2. Find the mean for dogs and the mean for cats.
3. Take the difference. **(Signal = $\text{mean}_{\text{dogs}}$ - $\text{mean}_{\text{cats}}$)**

Now we want to know if dogs are cuter than cats.

We could:

1. Get two samples: one of dogs and one of cats.
2. Find the mean for dogs and the mean for cats.
3. Take the difference. **(Signal = $\text{mean}_{\text{dogs}}$ - $\text{mean}_{\text{cats}}$)**

Testing the null hypothesis

Null hypothesis (H_0): There is **no difference** in the cuteness of dogs and cats.

Testing the null hypothesis

Null hypothesis (H_0): There is **no difference** in the cuteness of dogs and cats.

Alternative hypothesis (H_A): There **is** a difference in the cuteness of dogs and cats.

Testing the null hypothesis

Null hypothesis (H_0): There is **no difference** in the cuteness of dogs and cats.

Alternative hypothesis (H_A): There **is** a difference in the cuteness of dogs and cats.

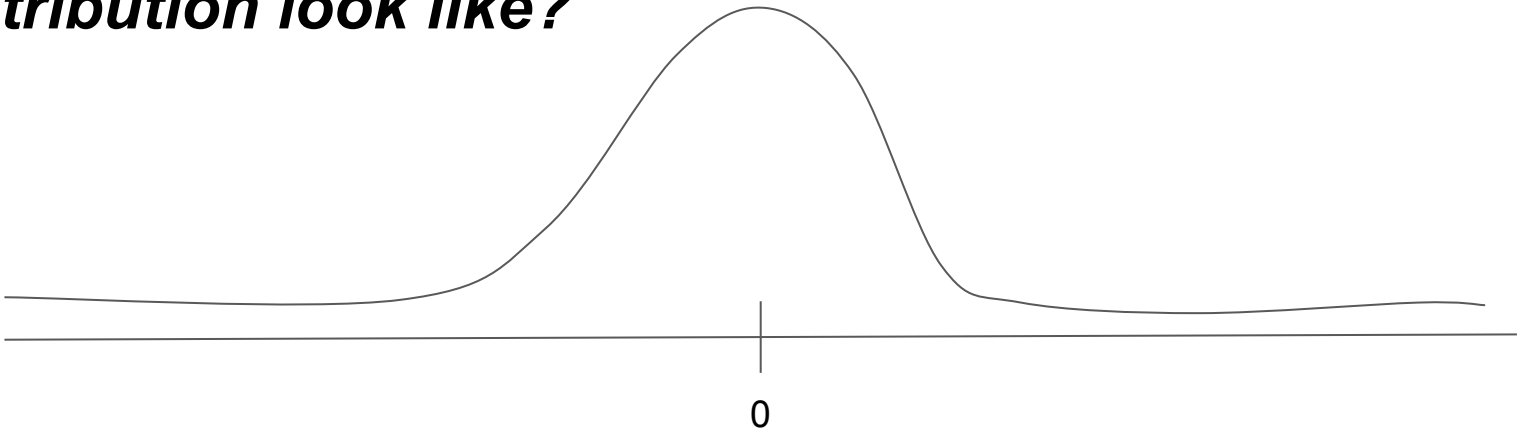
IF the null hypothesis were true, what would the sampling distribution look like?

Testing the null hypothesis

Null hypothesis (H_0): There is **no difference** in the cuteness of dogs and cats.

Alternative hypothesis (H_A): There **is** a difference in the cuteness of dogs and cats.

IF the null hypothesis were true, what would the sampling distribution look like?

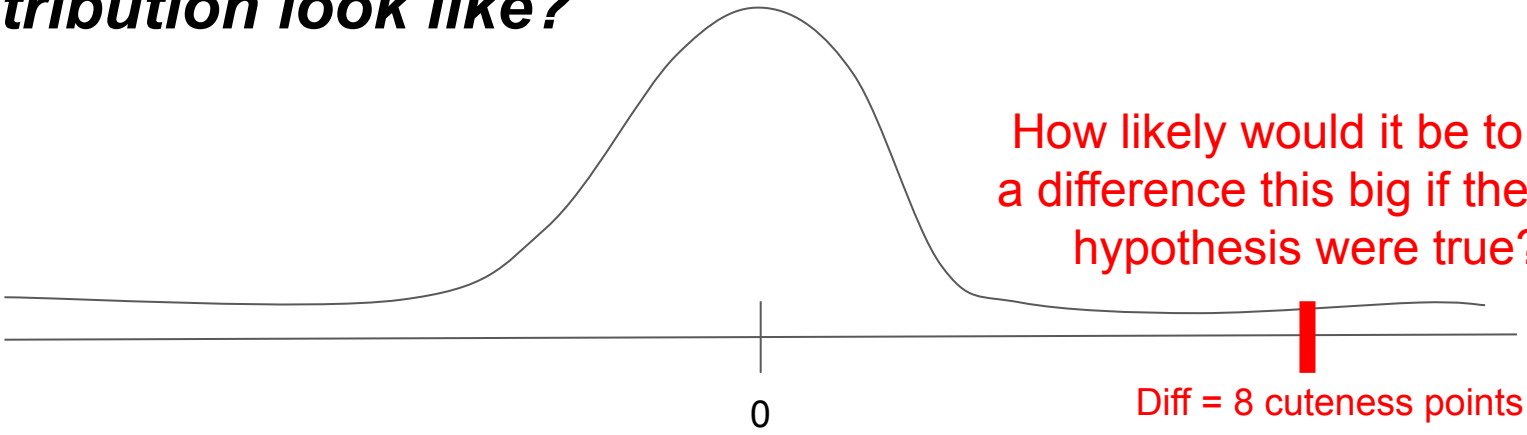


Testing the null hypothesis

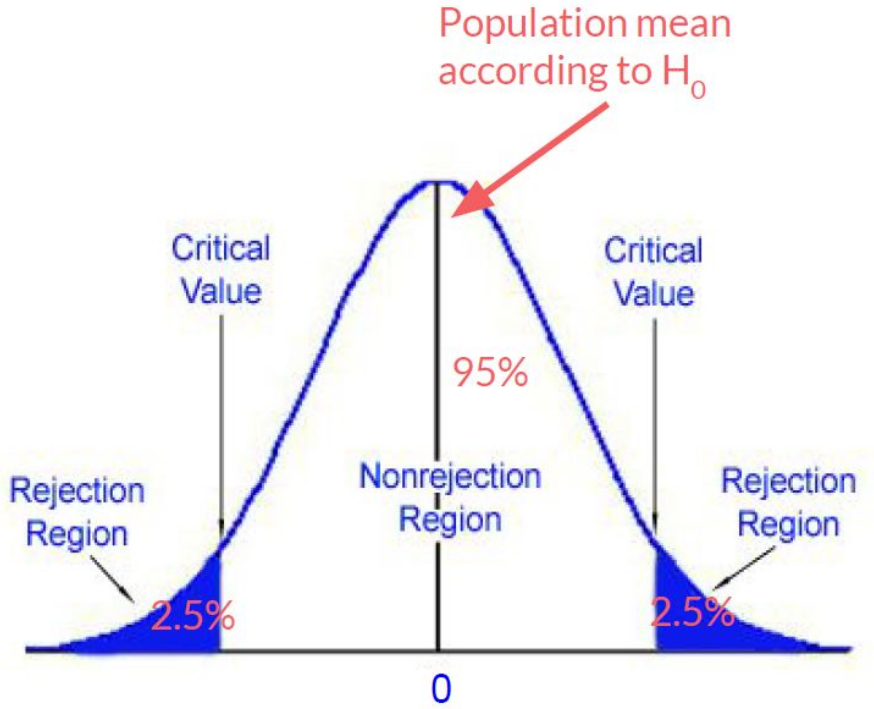
Null hypothesis (H_0): There is **no difference** in the cuteness of dogs and cats.

Alternative hypothesis (H_A): There **is** a difference in the cuteness of dogs and cats.

IF the null hypothesis were true, what would the sampling distribution look like?



Sampling distribution, with a standardized scale



(t-distribution, or Normal distribution)

Testing the null hypothesis

How likely would it be to get a difference this big if the null hypothesis were true?

If the probability < 0.05, we say it would be UNLIKELY.

→ We reject the null hypothesis.

Otherwise…

→ We FAIL to reject the null hypothesis.

Testing the null hypothesis

How likely would it be to get a difference this big if the null hypothesis were true?

If the probability < 0.05, we say it would be UNLIKELY.

→ We reject the null hypothesis.

→ We conclude there is a ***statistically significant*** difference in the cuteness of dogs and cats.

Otherwise…

→ We FAIL to reject the null hypothesis.

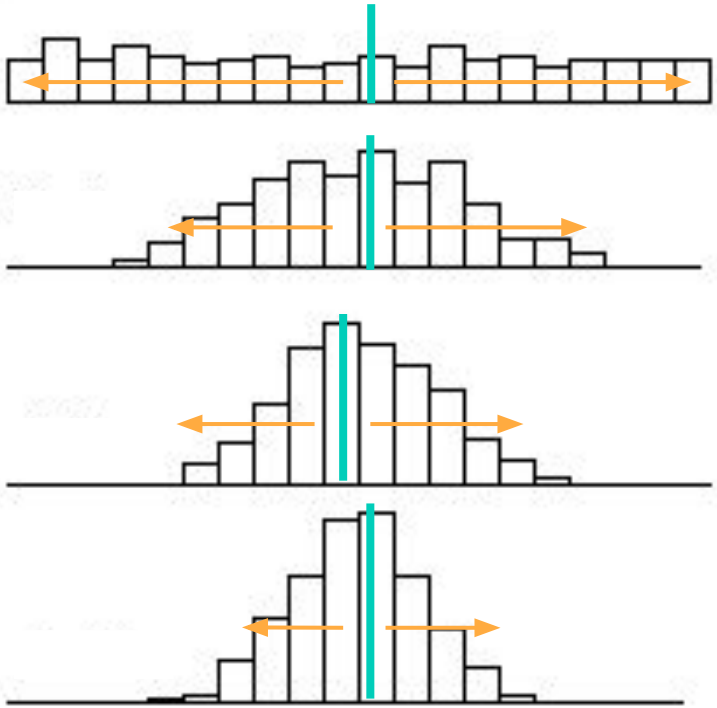
→ We cannot conclude there is any significant difference in the cuteness of dogs and cats.

Pros and cons of using $p<0.05$ as a cutoff:

Pros and cons of using p<0.05 as a cutoff:

Some recent challenges to the null hypothesis significance test:

- Arbitrariness of 95% confidence intervals
- The difference between significant and not significant is itself not statistically significant
- The null hypothesis that $\mu_1 = 0$ is frequently literally incredible. This becomes more and more apparent as datasets get larger and larger.



Lunch

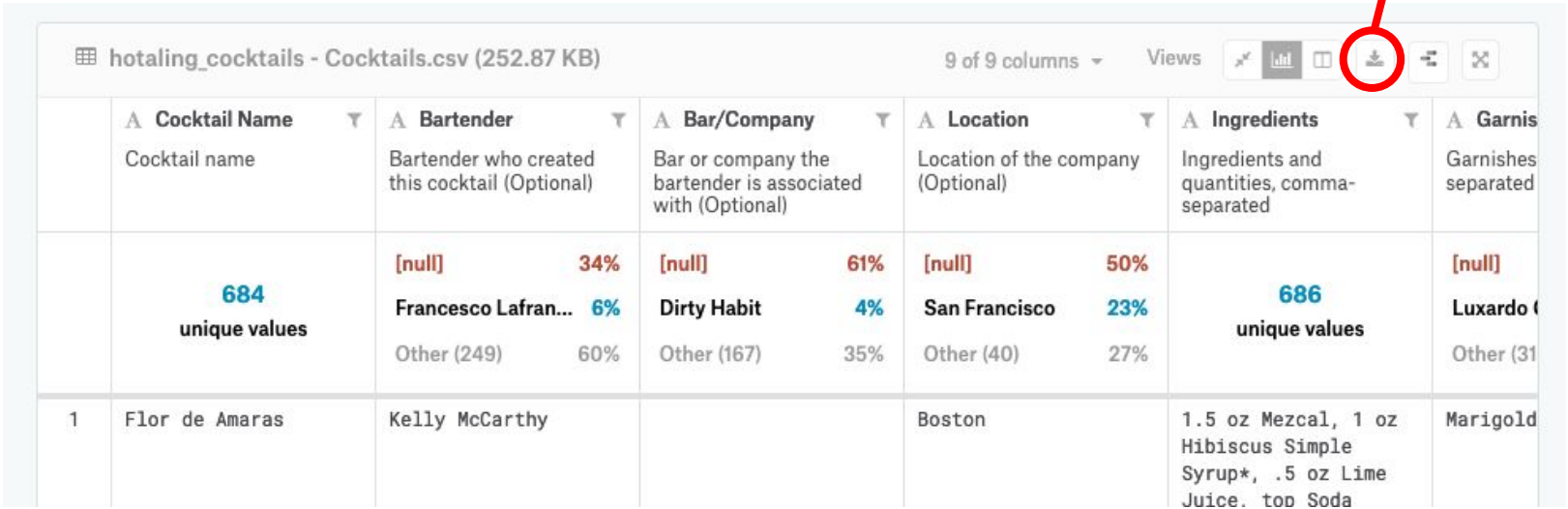
Managing projects

Goals for Project Management

- Practical:
  - ...organize programming files to help future you and others.
  - …organize yourself for 381
- Conceptual:
  - ...intentionally develop your style.
  - ...imagine stats projects as rivers, trees, or webs.

Get some data!

1. Go to www.kaggle.com/datasets
2. Make an account (sorry)
3. Use the search filter to download a .csv file < 2MB



Make a folder structure and project

1. First make a project folder with the name of your project
2. Inside your project folder include folders for:
    a. Raw data, Clean data
    b. Cleaning, Analysis
    c. Graphs, Logs
    d. Archive
3. Create a new Project in Stata in your project folder
4. Drag your folder structure into the project manager!

Convert you raw data to .dta

1. Move your .csv data to your Raw Data folder
2. Create a new .do file in your Cleaning folder
3. Add starter code and comment block
4. `import delimited “Raw Data/YOUR_CSV.csv”, varnames(1)`
5. `save “Clean Data/NICE_DATA.dta”, replace`
6. Drag your nice new data from the finder into your project manager folder tree

Analyze!

1. Generate summary statistics for a couple variables
2. Make 2 histograms for two interesting continuous variables
   a. `histogram varname`
3. Output each figure to your Plots folder
   a. `graph save "Graphs/filename", replace`
4. Make a scatter plot of the two variables and save to your graphs folder.
5. Play with the advanced stata graphics in the Day 2 folder!

TWIST

1. Add two different mistakes to your analysis .do file
2. Add a to-do item [be merciful and don’t make it too hard]
    - e.g., summarize a var, generate a new var, plot a graph
3. Compress your project folder, share it to a partner you haven’t worked with yet.
4. Open the file, find the errors, complete the to-do!

Reflect/Radiate

https://forms.gle/UwyPtY6kFEPgm47JA