Welcome to DSCI 101

Introduction to Data Science

Week 1-8 Review

- Exploratory Data Analysis (EDA)
  - understanding your data

- EDA goals and techniques
  - 5 key properties and 3 main aspects to explore
  - univariate, bivariate and multivariate analysis
  - feature engineering

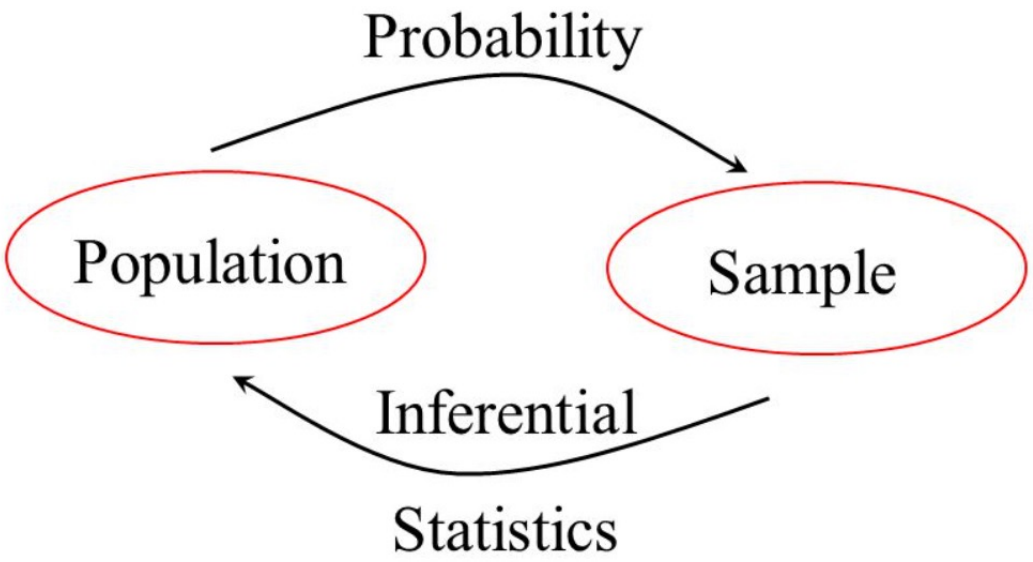
Week 9 Preview

- Introduction to Probability
  - random variable
  - probability distribution

- Statistical Inference
  - Important concepts: Population and parameters, Sample and statistics
  - Sampling distribution and empirical distribution
  - Fundamental Theorem of Statistics

- Quantify uncertainty with Bootstrap

Probability and Statistics



Random Event

- A random event
  - collection of outcome of some random process
  - sample space of all possible outcomes

- Set operations
  - complement: A and A^c, same as logical operator **negate**
  - union: $A \cup B$, same as logical operator **or**
  - intersection: $A \cap B$, same as logical operator **and**

- Examples: flip a coin, throw a die, etc.

Basic probability rules

1) General rule:
   - $0 \leq \mathbb{P}(A) \leq 1$ for any event A
   - in particular, $\mathbb{P}(\phi) = 0$ for empty set, and $\mathbb{P}(S) = 1$ for entire sample space

2) Addition rule: A or B
   - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
   - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if they are mutually exclusive (or disjoint)

3) Multiplication rule: A and B
   - $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B \mid A)$ (B conditioned on A) $= \mathbb{P}(B) \times \mathbb{P}(A \mid B)$
   - $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$ if they are independent

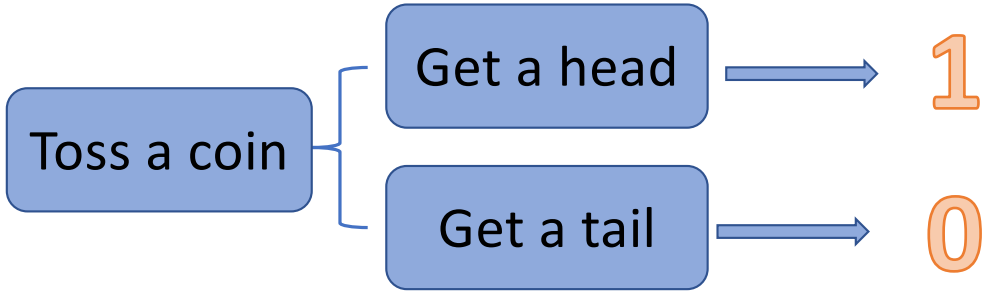
Review of probability sampling

- Knowing what's your population
- Knowing the probability of selecting any subgroup in the population

- Remember df. sample()?
- Difference between sampling with and without replacement
  - real world population is finite, so it's always without replacement
  - but when population is huge compared to sample size, with replacement is a close enough approximation of reality

Random variable

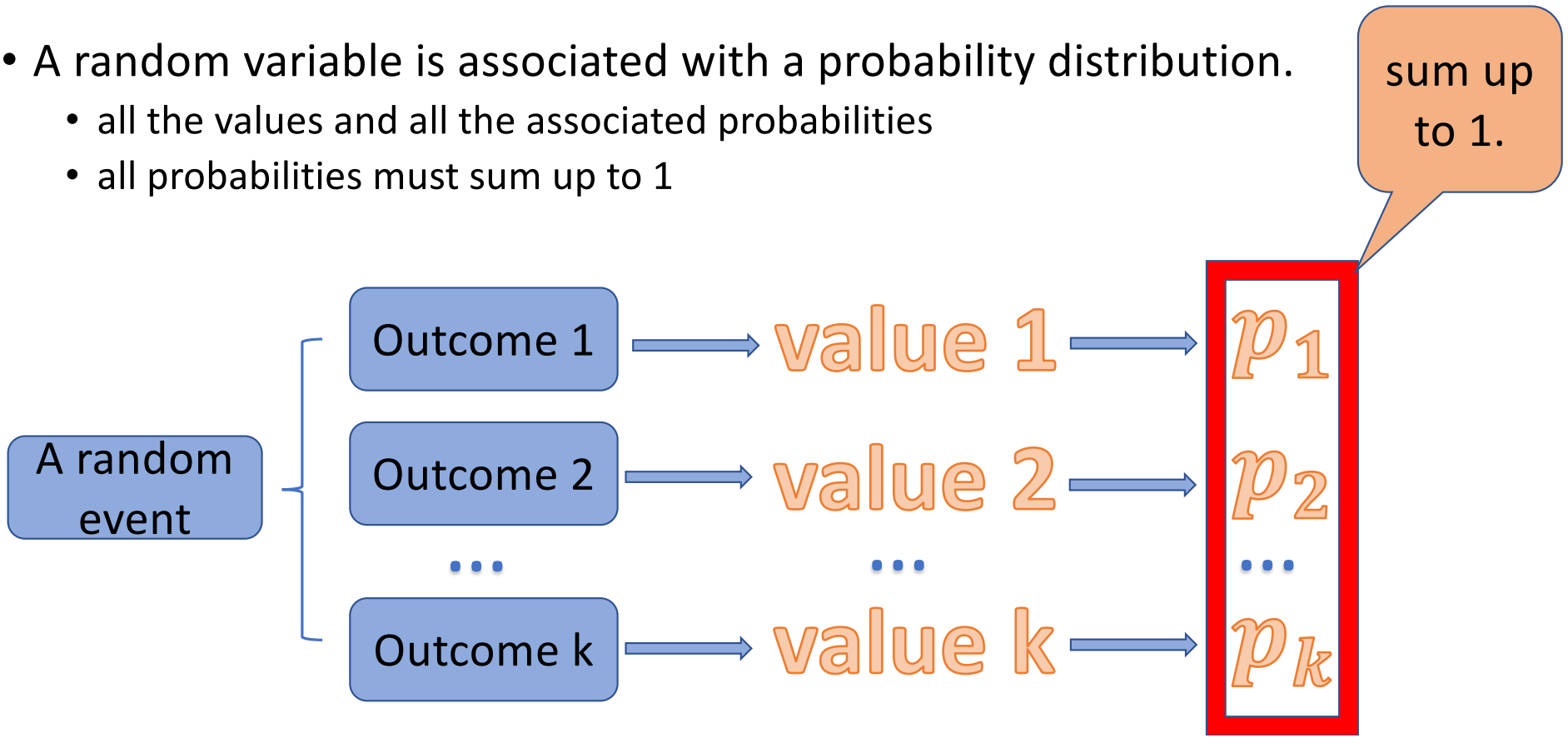
- A random variable is a variable whose values depend on some random event outcome.
- Mathematically, a random variable is a **function that maps** a random event outcome to a real number.



- Toss a coin is NOT a random variable, but this mapping is.

Probability distributions

- A random variable is associated with a probability distribution.
  - all the values and all the associated probabilities
  - all probabilities must sum up to 1



Types of random variables and distributions

- Same as discrete vs. continuous numerical variable types
- Discrete: you can count the different values of the r.v.
- Continuous: you can not!

- focus almost exclusively on discrete distributions, and one continuous

- more in a calculus-based probability course!

Common discrete distributions

- Bernoulli(p)
  - A r.v. takes value 1 with probability p and value 0 with probability 1-p
- Binomial (n, p)
  - A r.v. which is sum of n independent Bernoulli(p) r.v.'s
- Uniform on a finite set of k different values and equal p=1/k
  - For example: roll a fair die

Bernoulli(p) = Binomial (1, p)

Numpy.random

- Simulation allows you to generate random samples from a specified probability distribution
- ***np.random.dist_name(dist_parameter_values, sample_size)***

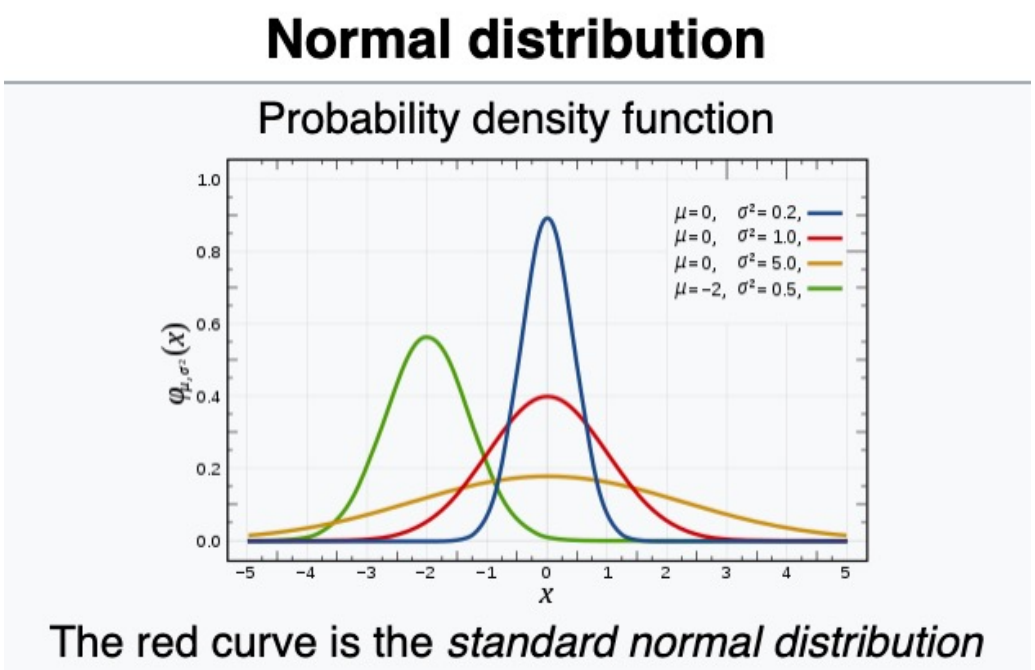
- Examples:
  - ***np.random.binomial(n=1, p=0.5, size=100)***
  - ***np.random.binomial(n=100, p=0.2, size=1)***
  - ***np.random.randint(low=1, high=11, size=3)***

Normal distribution

- A continuous distribution with two parameters: mean and variance
- Possible values: all real numbers
- Symmetrical and bell-shaped

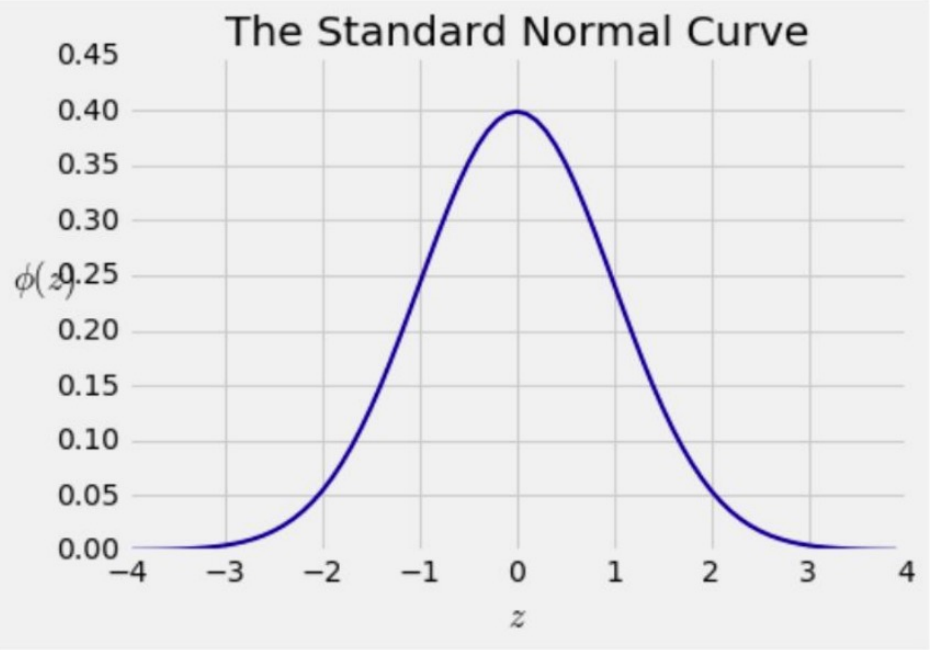
- Mean μ:
  - the center / symmetrical line

- Variance (sd) σ^2:
  - the spread out



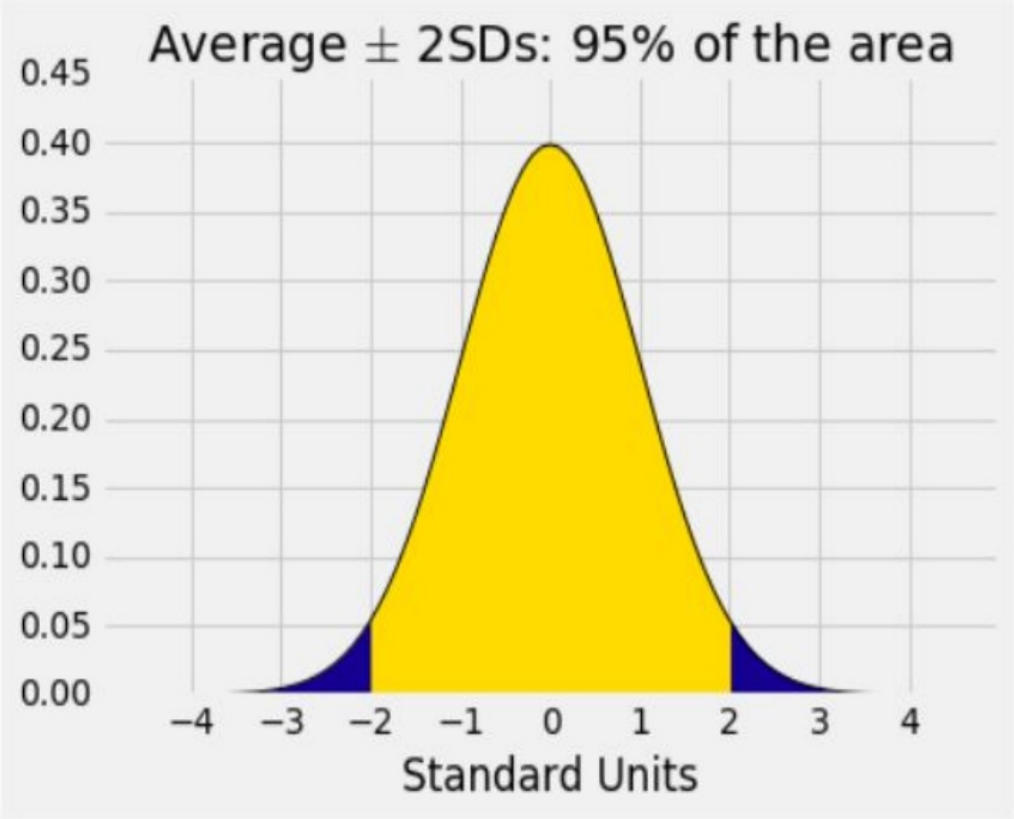
The standard normal curve

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad -\infty < z < \infty$$



- A beautiful formula (that we won't use at all)
- A special case of normal:
  - $\mu = 0$
  - $\sigma^2 = 1$
- Any normal distributions can be converted to standard normal
  - also known as z-score

The 2 SD rule



- Some convention:
  - yellow = usual values
  - blue = extreme values (two tails)
- Area represents probability!
  - middle area with $\mathbb{P}$ = 95%
  - tail areas with $\mathbb{P}$ = 5%
- Applies to any normal distributions!

Birthday Problem

- The **birthday problem** asks for the probability that, in a set of *n* randomly chosen people, at least two will share a birthday.
  - what is your guess?

- How do we approach this problem using computer simulation?
  - simulate a random event repeatedly (a large number of times)
  - use long term proportion as approx. probability
  - this is called Monte Carlo simulation

Inferential Statistics: important concepts

- Population distribution
  - a census!
- Parameters: a quantity associated with population distribution
  - you will know this if you have a census!
- Random probability sample
  - to make inference about the population

Empirical distribution

- “Empirical”: based on the observations of your random sample

- Empirical distribution:
  - assume sample size of N observations (number of rows)
  - each observation has equal probability 1/N

- Bar plot and histogram of sample data
  - represent the empirical distribution
  - count values for each category or each bin
  - each observation counts as 1/N!

Sampling distribution

- Statistic: a quantity associated with your random sample
  - anything you can calculate based on your data
  - sample mean, sample median, sample SD, sample IQR…
- Sampling distribution of a statistic
  - different values and associated probabilities based on ALL possible random samples from the population
- Simulate the sampling distribution
  - repeatedly draw random samples – doable with simulation!

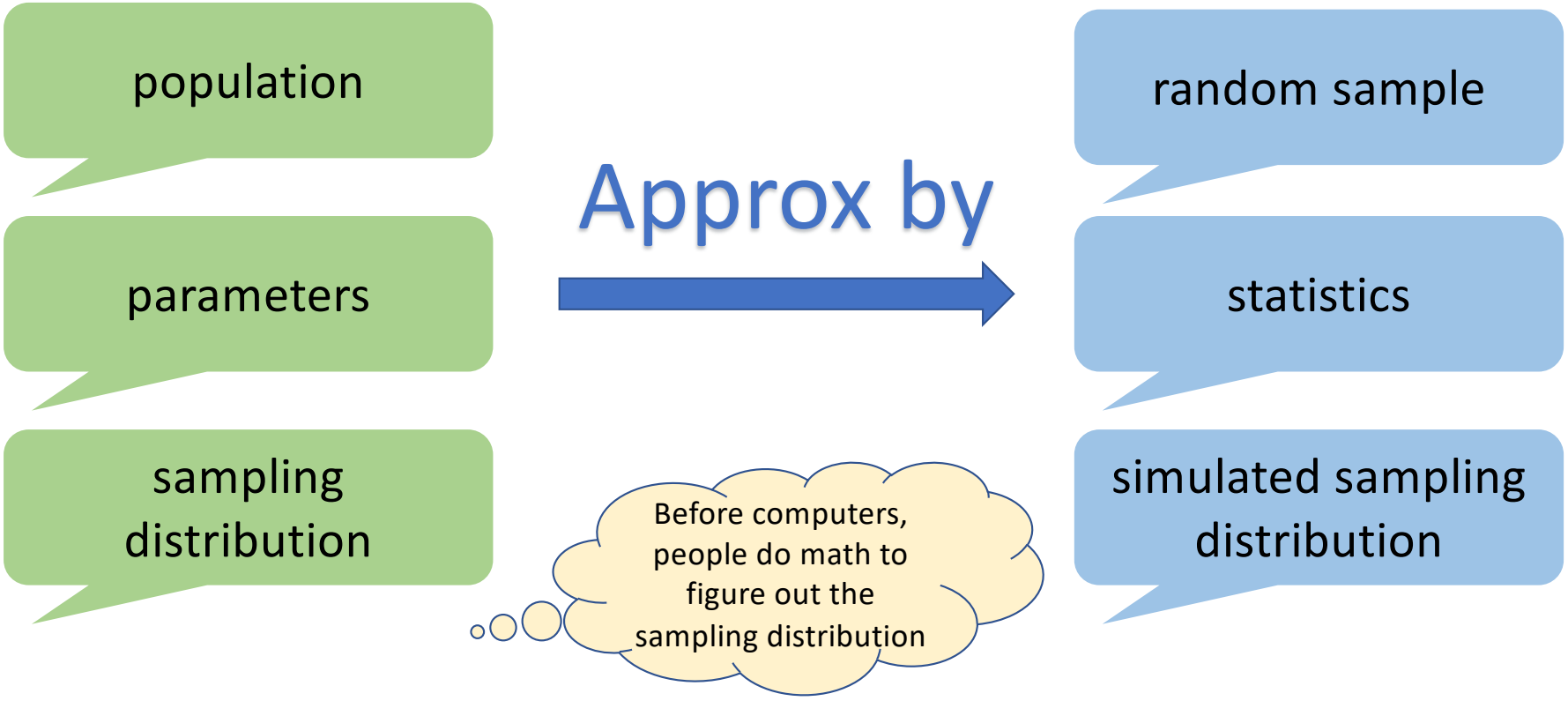
Why do we care?

- A statistic can approximate the population parameter
  - use sample mean to approximate population mean

- However, we know if we get a different sample, the sample mean could have come out different, but how different???
  - need to **quantify the uncertainty** → sampling distribution

- Quantify the uncertainty using simulation → simulate the sampling distribution

How to make statistical inference?



Simulate a statistic

- Computer is good at doing repetitive tasks!
- Figure out code to compute the statistic based on one sample
- Repeat: draw a random sample –> compute the statistics

- After enough repetitions, you will have
  - the (approx.) **sampling distribution** of that statistic
  - everything you need to know about that statistic

Simulate a statistic – pseudo code

- For a fixed sample size N:
    1. Draw a random sample of size N
    2. Calculate the statistic of interest based on one random sample

- Repeat 1&2 many many times… (A For statement in Python)

- Visualize: histogram of the simulated statistic
    - the sampling distribution of your statistic!!! (with large enough repetition)

Where to get another sample?

- What did you notice about the simulation we just did?
- We know the population distribution!

- But in the real world, we don't…….
- To collect another sample data requires time and money…lots of them for many samples

- Stuck?

The Bootstrap

- A technique to generate more random samples from one original sample.
- Huh???

- Theoretical justification: the Fundamental Theorem of Statistics
  - if your sample is large and random
  - the empirical distribution of your data $\approx$ the population distribution

- So we sample from the empirical distribution!

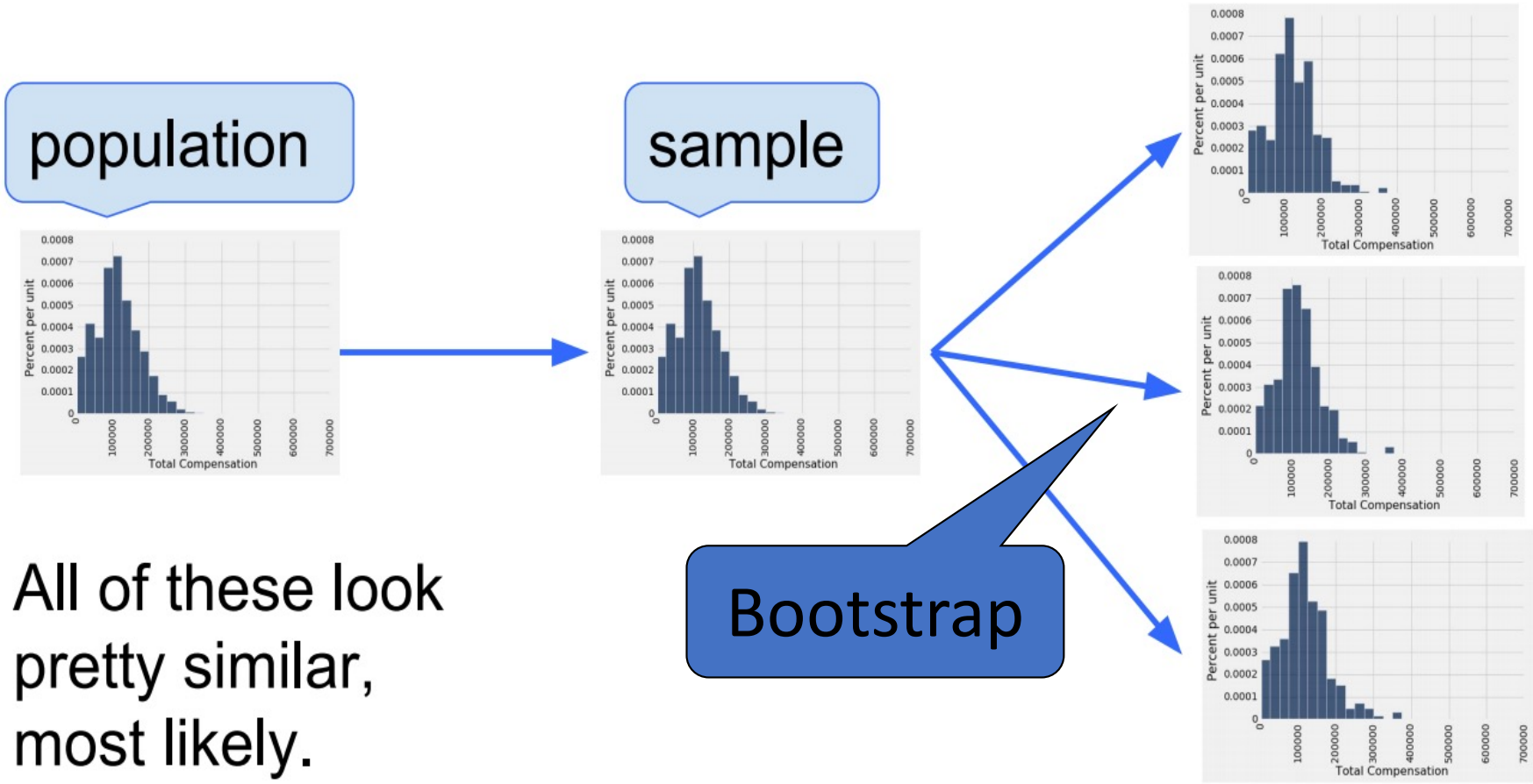
How???

- Remember what is the empirical distribution of your data?

- From the original sample of N observations:
  - to generate a **bootstrap resample**
  - draw **N (the same sample size)** observations at random **with replacement**

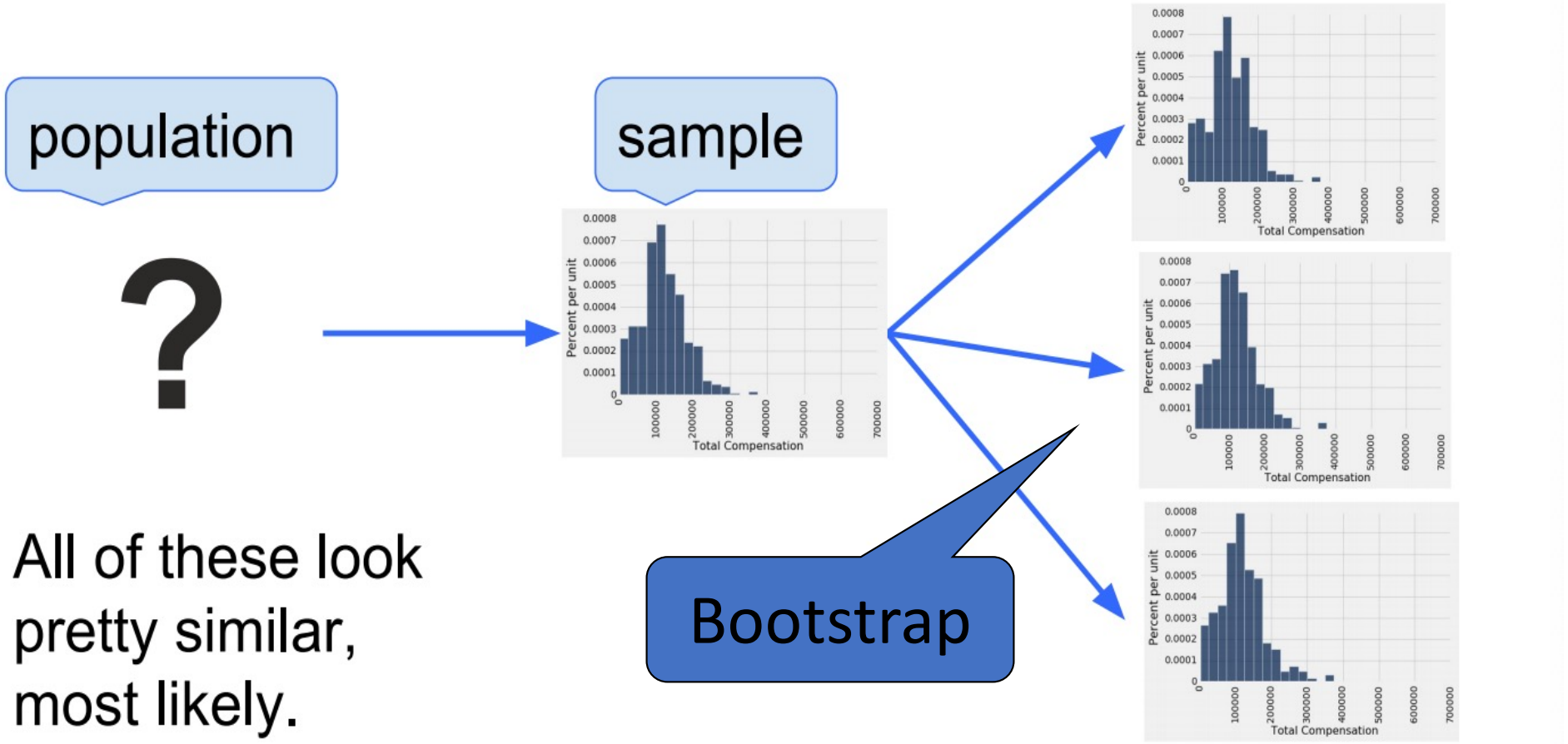
- If you are confused, think about:
  - how do you sample from a discrete uniform distribution, like a die?
  - what happens if you draw N observation without replacement?

Illustration



All of these look pretty similar, most likely.

Real world


population
?
sample
Percent per unit
Total Compensation
Bootstrap
All of these look pretty similar, most likely.

When Not to use Bootstrap

- When your original sample is NOT large and random!
  - very small sample size
  - highly dependent samples

- When the dataset you have is poor in quality,
  - not much you can do with it…..

"Garbage in, garbage out"

Your analysis is as good as your data.

Confidence Interval

- Interval estimate of the population parameter

- 90% is called the confidence level
  - “middle” 90% of the sampling distribution
  - higher confidence → wider intervals (trade-off)

- The confidence is in the process that generated the interval
  - it generates a “good” interval about 90% of the times

Take away message

- Probability is fun!
  - using simulation to figure out probability is more fun!
- Statistical inference is all about quantifying uncertainty
  - confidence intervals!!!
- Figuring out sampling distribution mathematically is super hard!
  - thanks to bootstrap and simulation