CS57300
PURDUE UNIVERSITY

JANUARY 17, 2019

DATA MINING

POPULATIONS AND SAMPLES

ELEMENTARY UNITS, POPULATIONS, AND SAMPLES

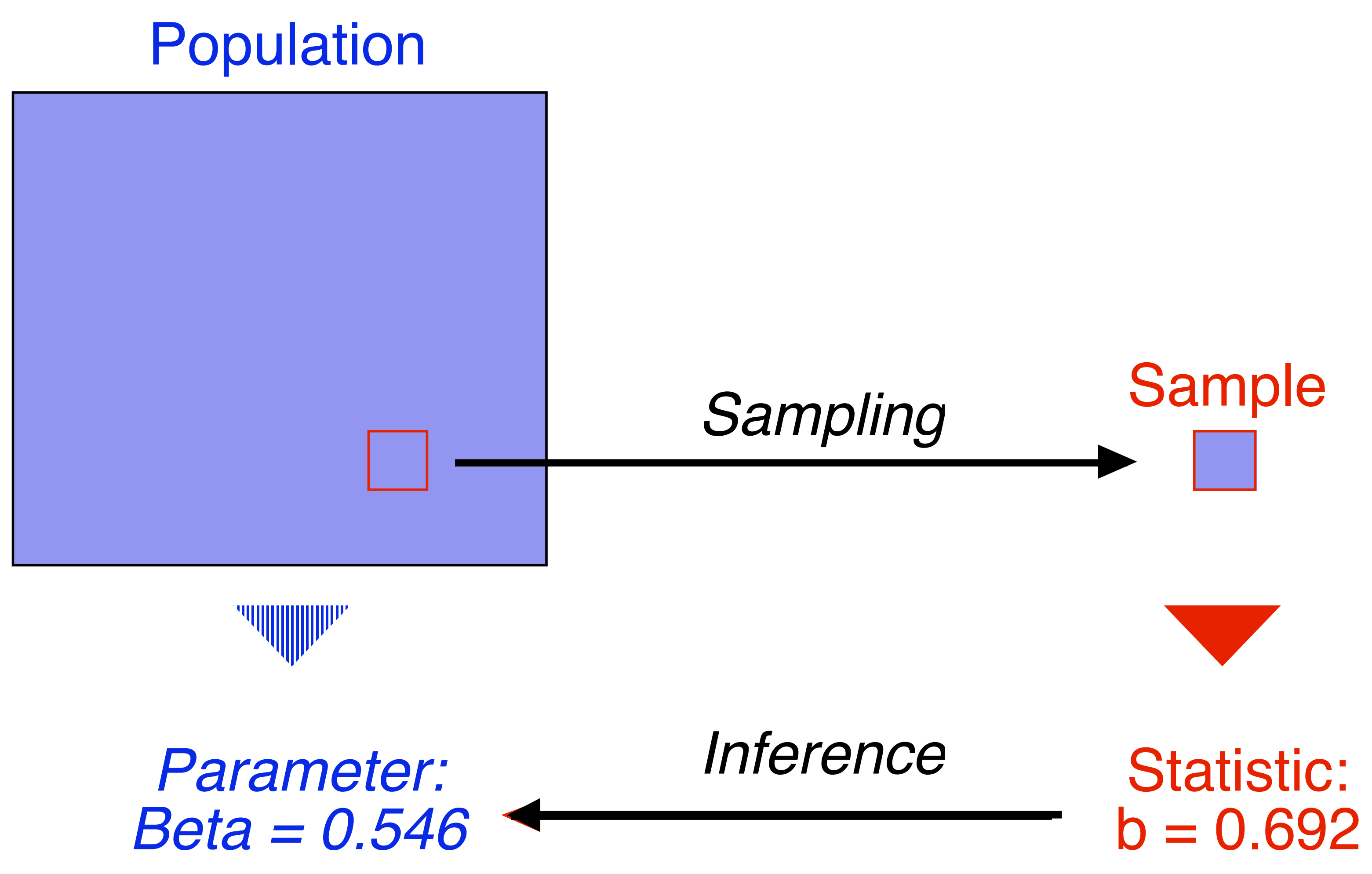
- Elementary units:
    - Entities (e.g., persons, objects, events) that meet a set of specified criteria
    - Example: A person who has purchased something from Walmart in the past month
- Population:
    - Aggregate of elementary units (i.e, all entities of interest)
- Sample:
    - Sub-group of the population

SAMPLING

- Reasons to sample
  - Obtaining the entire set of data of interest is too expensive or time consuming
  - Processing the entire set of data of interest is too expensive or time consuming
- Sampling is the main technique employed for data selection

USE SAMPLES FOR ESTIMATION

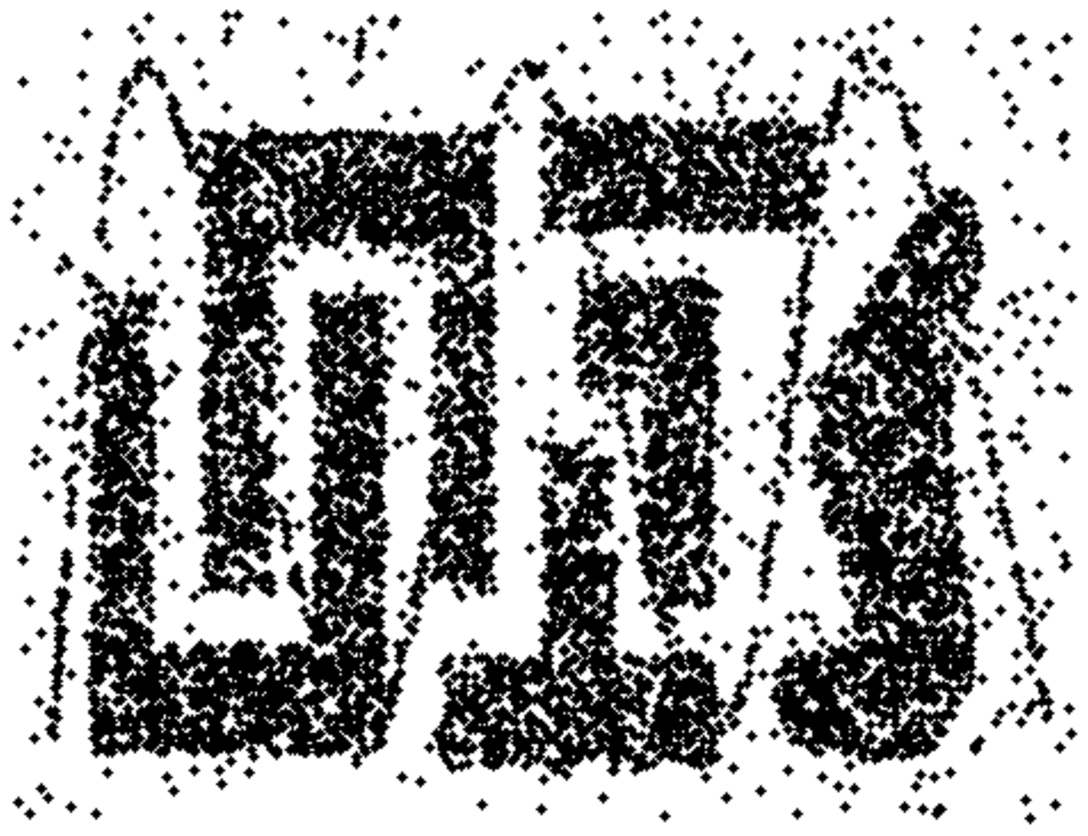
- In data mining we often work with a sample of data from the population of interest
- If we had the population we could calculate the properties of interest
- Sample serves as a reference group for **estimating** characteristics about the population and drawing conclusions



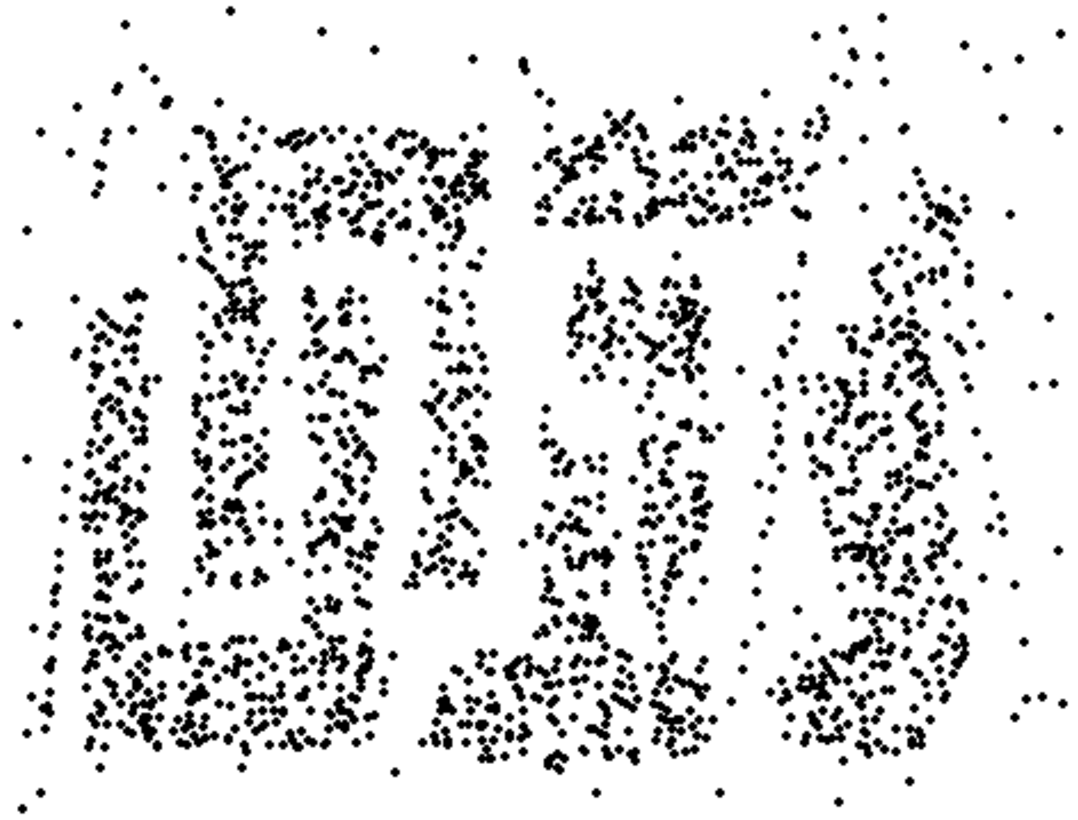
PRINCIPLE FOR EFFECTIVE SAMPLING

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, if the sample is **representative**
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

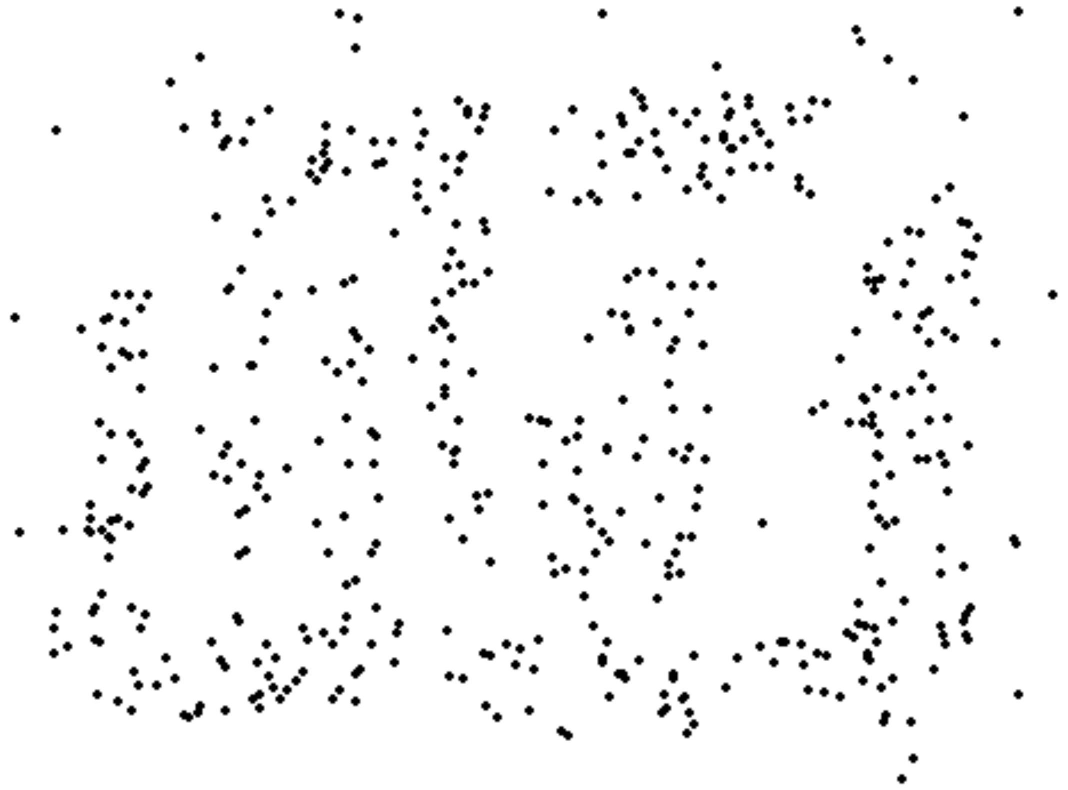
SAMPLE SIZE



8000 Points



2000 Points



500 Points

TYPES OF PROBABILITY SAMPLING

Covariance between samples with replacement is 0

- Simple random sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement
    - As each item is selected, it is removed from the population
  - Sampling with replacement
    - Items are not removed from the population as they are selected for the sample; the same item can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

SAMPLING METHODS

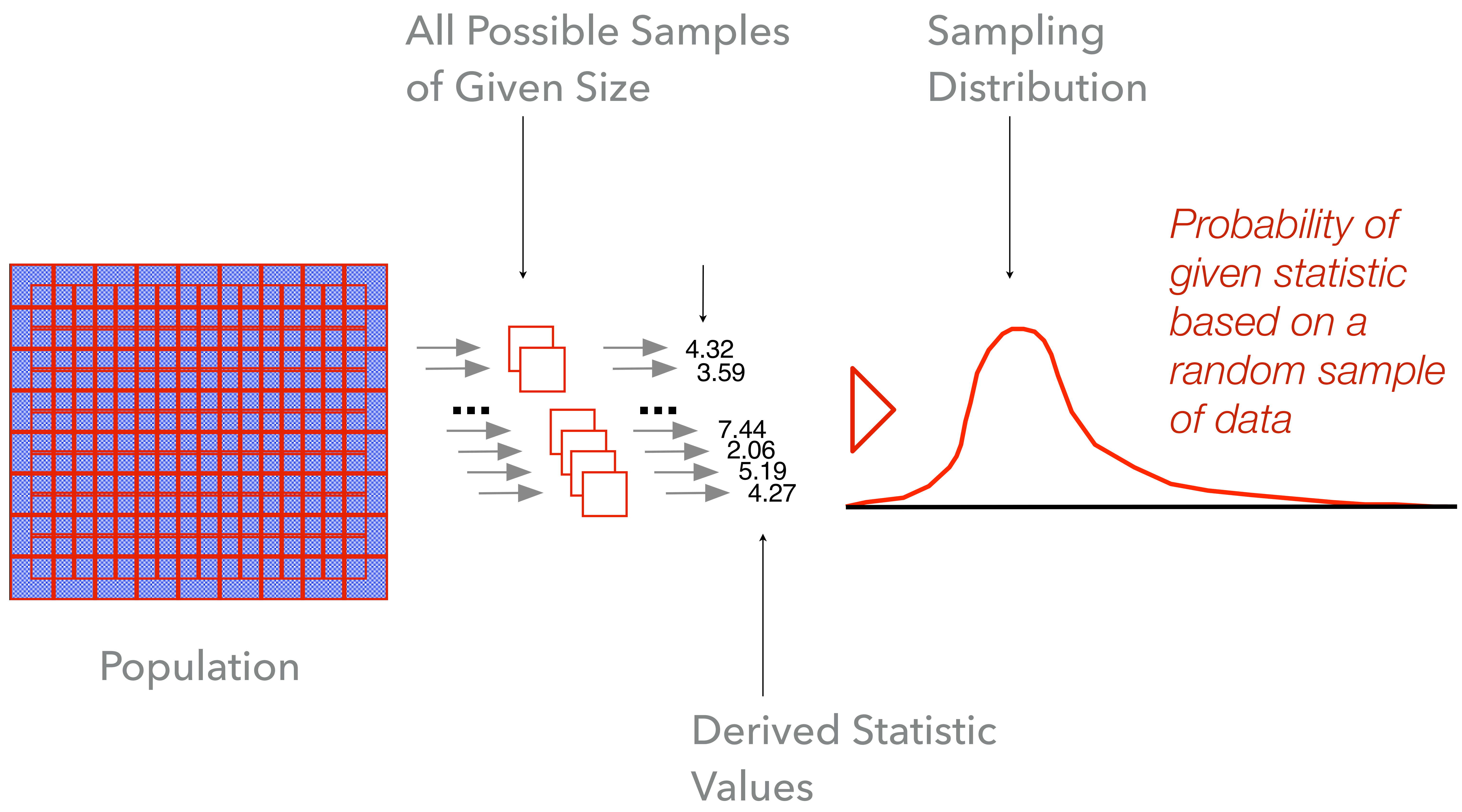
PROBABILITY SAMPLING

1. SIMPLE RANDOM SAMPLING
2. SYSTEMATIC SAMPLING
3. STRATIFIED SAMPLING
4. CLUSTER SAMPLING
5. MULTI-STAGE SAMPLING

NON-PROBABILITY SAMPLING

1. CONVENIENCE SAMPLING
2. QUOTA SAMPLING
3. SNOWBALL SAMPLING
4. JUDGEMENT SAMPLING

SAMPLING DISTRIBUTIONS



STATISTICAL INFERENCE

STATISTICAL INFERENCE

- Infer properties of an unknown distribution with sample data generated from that distribution
- Parameter estimation
  - Infer the value of a population parameter based on a sample statistic (e.g., estimate the mean)
- Hypothesis testing
  - Infer the answer to a question about a population parameter based on a sample statistic (e.g., is the mean non-zero?)

PARAMETER ESTIMATION

- Infer the value of population parameters (θ) from data
- θ can take values in the parameter space Θ
- Frequentist approach
  - Population parameters are fixed but unknown
  - Data is a random sample drawn from population
  - Use maximum likelihood estimation (MLE)
- Bayesian approach
  - Parameters are random variables with a distribution of possible values
  - Data is fixed and known, provides evidence for different parameter values
  - Use maximum aposteriori estimation (MAP)

https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

- Suppose we have a set of data $X = \{x_i\}_{i=1}^{N}$ independently drawn from the population
- The maximum likelihood estimation finds the parameter values that maximize the likelihood of observing the data

$$\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta} P(X|\theta) \\
&= \arg\max_{\theta} \prod_i P(x_i|\theta) \\
&= \arg\max_{\theta} \log \prod_i P(x_i|\theta) \\
&= \arg\max_{\theta} \sum_i \log P(x_i|\theta)
\end{aligned}$$

MAXIMUM A-POSTERIORI ESTIMATION (MAP)

- Suppose we have a set of data $X = \{x_i\}_{i=1}^{N}$ independently drawn from the population, and the prior distribution for the parameter is $P(\theta)$
- The maximum a-posteriori estimation finds the mode of the posterior distribution of the parameters

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} P(X|\theta)P(\theta) \\ &= \arg\max_{\theta} \log P(X|\theta)P(\theta) \\ &= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta)P(\theta) \\ &= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)P(\theta)\end{aligned}$$

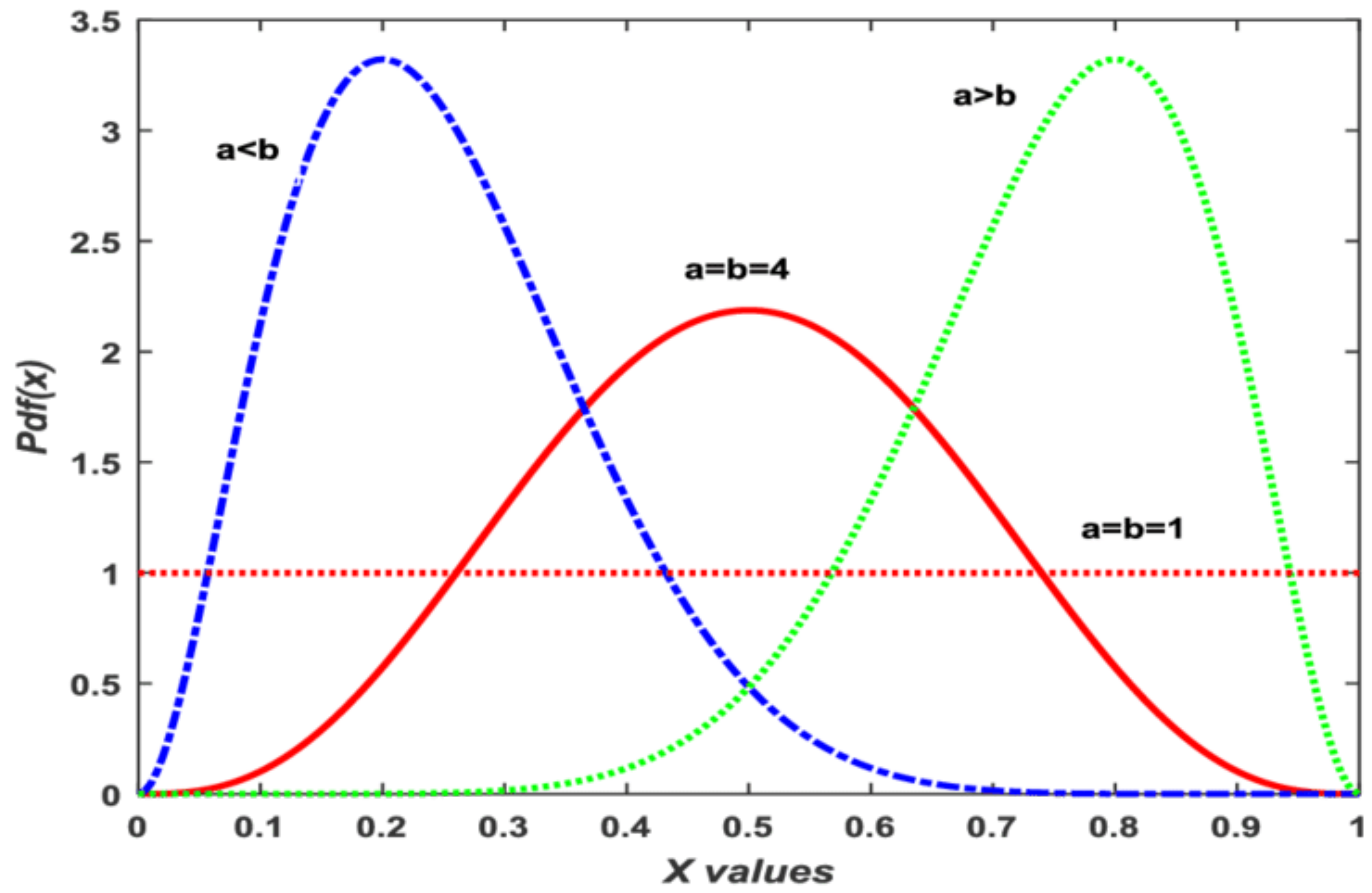
MLE VS. MAP EXAMPLE

- Flip a coin for *N* times and observe *n* heads; what's the probability of seeing the head if tossing the coin once?
- Likelihood of observing the data: $P(D\,|\,\theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n}$
  - The number of heads observed follows a binomial distribution
- Maximum likelihood estimation:

$$\theta_{MLE} = \text{argmax}_{\theta} P(D\,|\,\theta) = \frac{n}{N}$$

https://towardsdatascience.com/parameter-inference-maximum-aposteriori-estimate-49f3cd98267a

MLE VS. MAP EXAMPLE



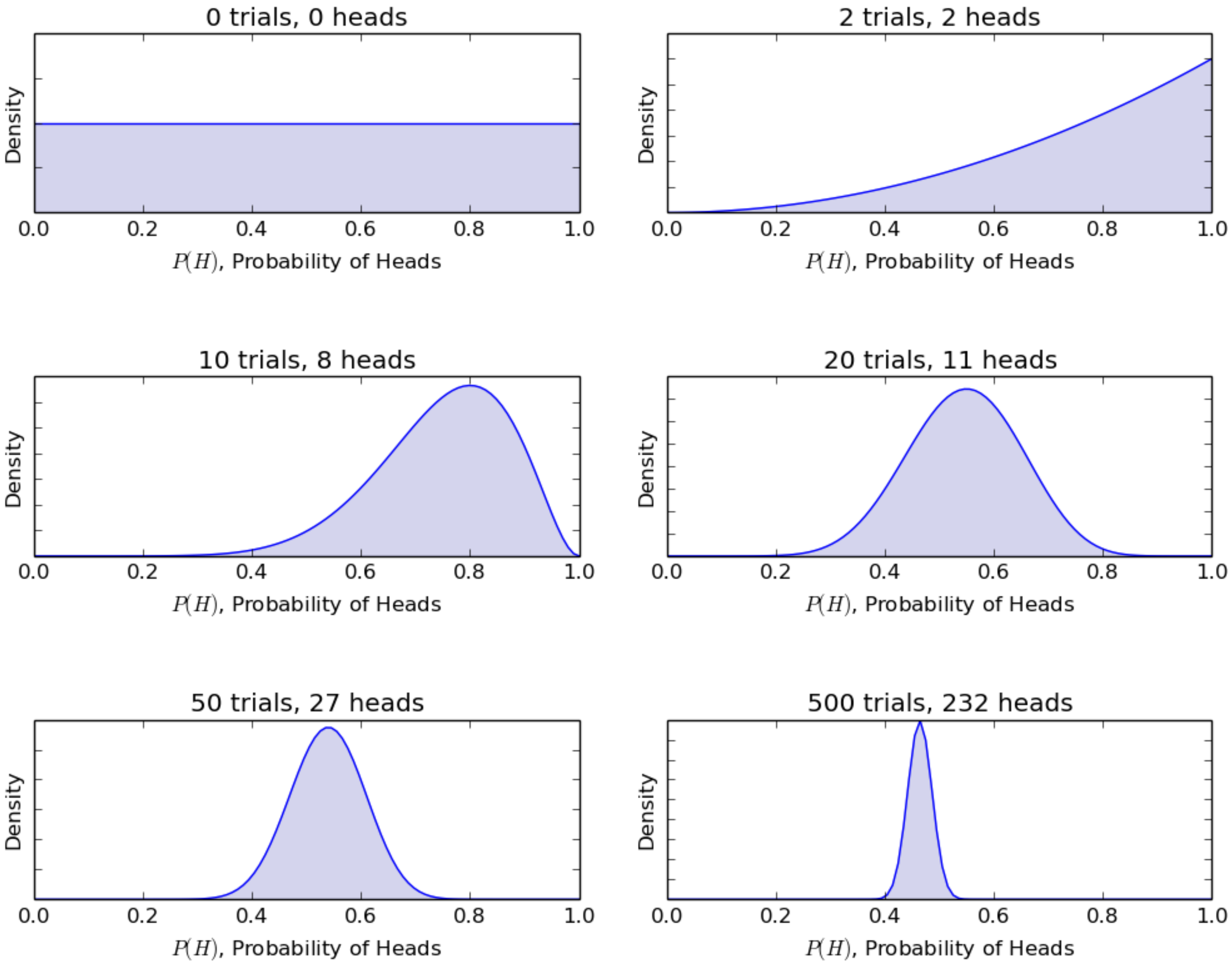
- Maximum a-posteriori estimation:
  - Suppose the prior is a Beta distribution

$$P(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \sim Beta(a,b), \text{where } B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$$

  - Then, the posterior is:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{\binom{N}{n}(\theta^{a+n-1}(1-\theta)^{b+N-n-1}/B(a,b))}{\int_0^1 \binom{N}{n}(\theta^{a+n-1}(1-\theta)^{b+N-n-1}/B(a,b))d\theta}$$

$$\sim Beta(a+n, b+N-n)$$


0 trials, 0 heads
2 trials, 2 heads
10 trials, 8 heads
20 trials, 11 heads
50 trials, 27 heads
500 trials, 232 heads
Density
0.0
0.2
0.4
0.6
0.8
1.0
$P(H)$, Probability of Heads

MLE VS. MAP EXAMPLE

- Maximum a-posteriori estimation:

$$\theta_{MAP} = \text{argmax}_{\theta} P(\theta \mid D)$$
$$= \text{argmax}_{\theta} Beta(a+n, b+N-n)$$
$$= \frac{a+n-1}{a+b+N-2}$$

- Notice that in this example, the posterior distribution is in the same probability distribution family as the prior distribution.
  - The prior and posterior are called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood function.
  - The beta distribution is a conjugate prior to the binomial likelihood.

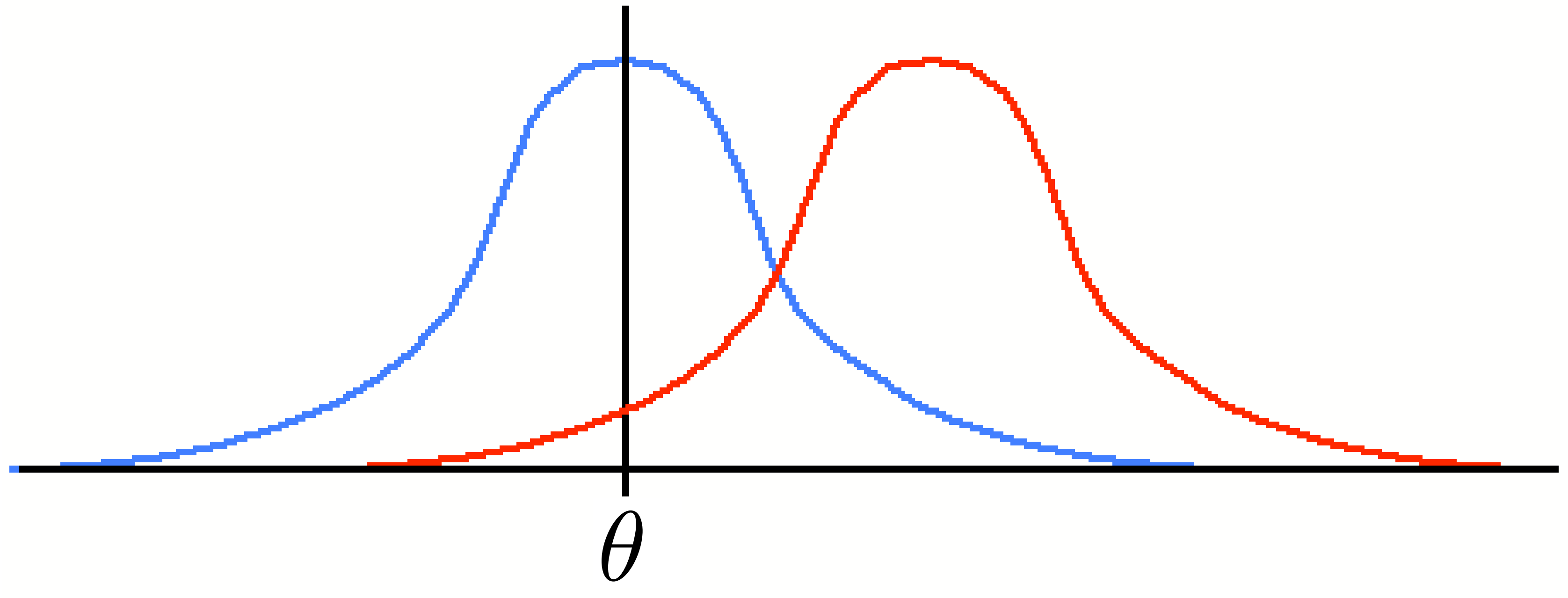
PROPERTIES OF ESTIMATORS

PROPERTIES OF ESTIMATORS

- Let $\hat{\theta}$ be an estimate for a population parameter θ
- Using different samples *D* will result in different estimates $\hat{\theta}_D$
- Thus $\hat{\theta}$ is a random variable with a distribution, mean, and variance
  - We can evaluate the quality of an estimator for θ based on the properties of the sampling distribution of $\hat{\theta}$

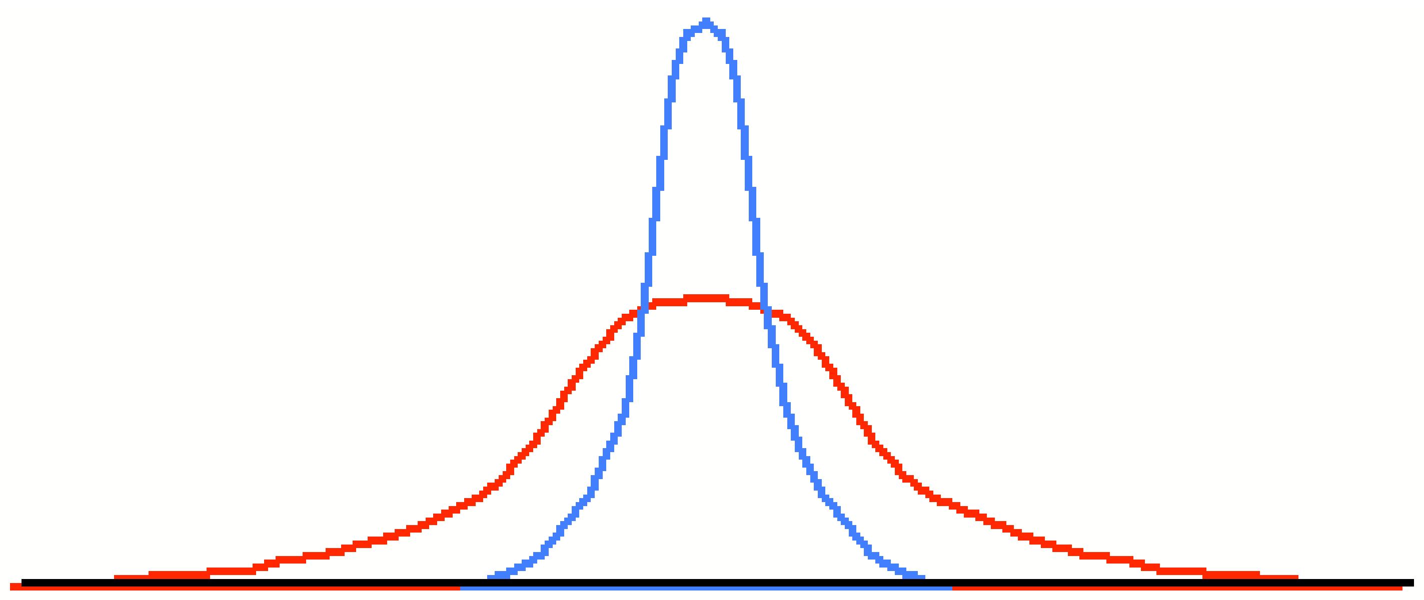
BIAS

- The best estimators produce values that center around the population parameter
- The **bias** of an estimator is defined as: $Bias(\hat{\theta}) = \underbrace{E[\hat{\theta}]}_{\text{Average estimated parameter}} - \underbrace{\theta}_{\text{True parameter in popul.}}$
- An estimator is unbiased if: $E[\hat{\theta}] - \theta = 0$



VARIANCE

- The best estimators produce values that differ only slightly from the population parameter
- The **variance** of an estimator is defined as: $Var(\hat{\theta}) = E[(\underbrace{\hat{\theta}}_{\text{Single parameter estimate}} - \underbrace{E[\hat{\theta}]}_{\text{Average estimated parameter}})^2]$
- Measures how sensitive the estimator is to different datasets
- Unbiased estimators with minimum variance are called *best unbiased estimators*



EXAMPLE

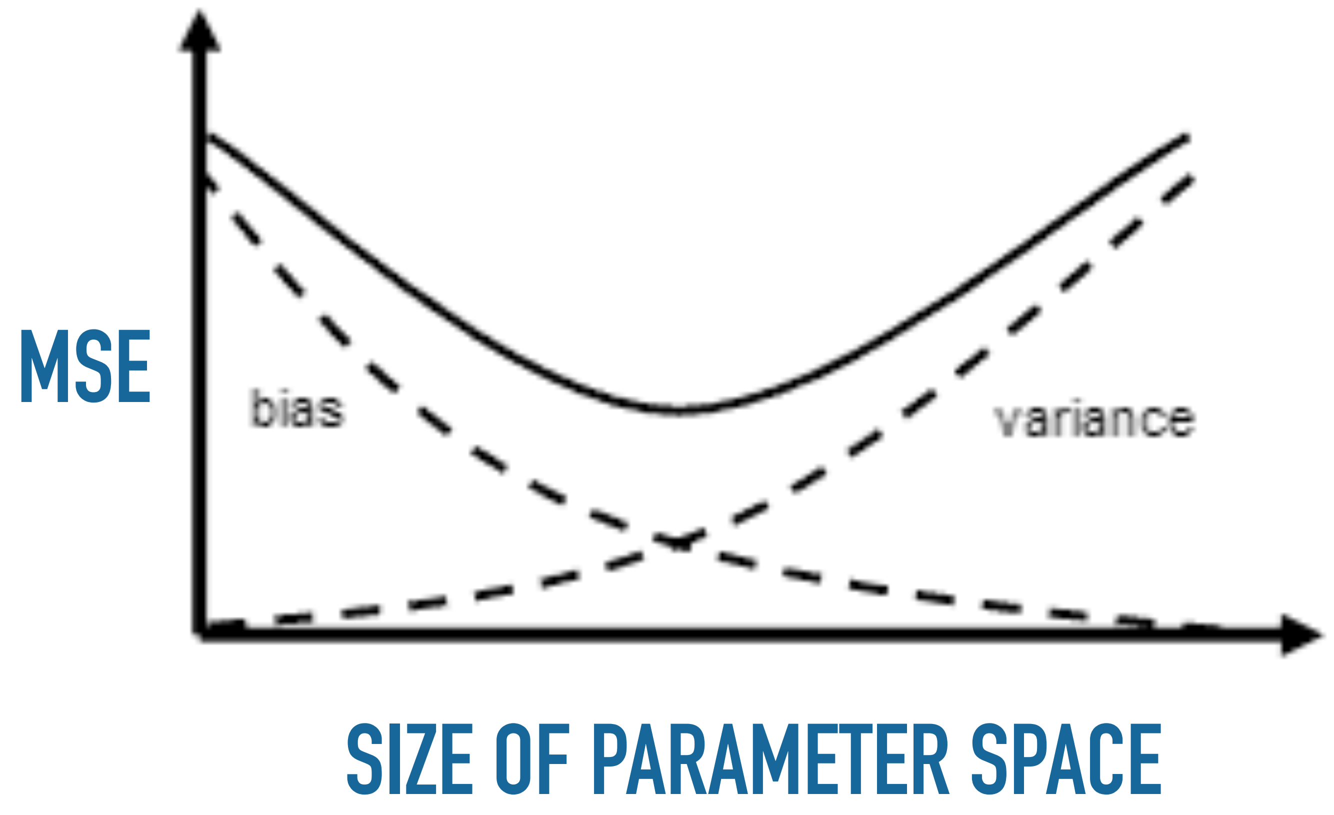
- Ignore data and declare that: $\hat{\theta} = 1.0$
- Estimate will not depend on data, thus: $Var(\hat{\theta}) = 0$
- However, in most cases this estimator will have a large bias (non-zero)

BIAS-VARIANCE TRADEOFF

- The mean-squared error (MSE) of $\hat{\theta}$ is:

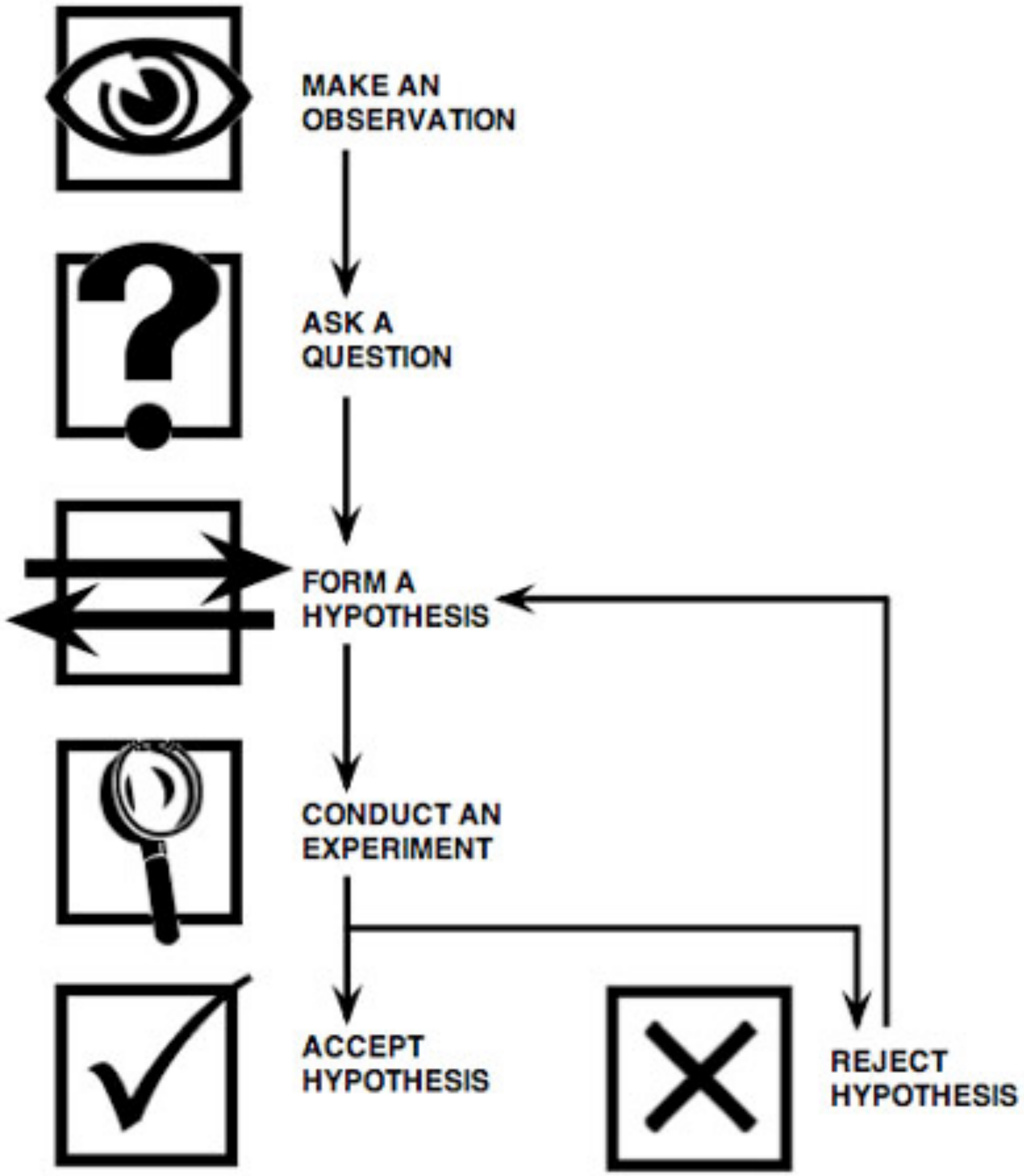
$$\begin{aligned} E[(\hat{\theta}-\theta)^2] &= E[(\hat{\theta}-E[\hat{\theta}]+E[\hat{\theta}]-\theta)^2] \\ &= \underbrace{(E[\hat{\theta}]-\theta)^2}_{\textbf{bias}} + \underbrace{E[(\hat{\theta}-E[\hat{\theta}])^2]}_{\textbf{variance}} \end{aligned}$$

- MSE measures systematic bias and random variance between estimate and population value
- Tradeoff: reducing bias tends to increase variance and vice versa



HYPOTHESIS TESTING

SCIENTIFIC METHOD



TYPES OF HYPOTHESES

Broad categories

- **Descriptive**: propositions that describe a characteristic of an object
- **Relational**: propositions that describe relationship between 2+ variables
- **Causal**: propositions that describe the effect of one variable on another

Specific characteristics

- **Non-directional**: an differential outcome is anticipated but the specific nature of it is not known (e.g., the tuning parameter will affect algorithm performance)
- **Directional**: a specific outcome is anticipated (e.g., the use of pruning will increase accuracy of models compared to no pruning)

Descriptive Hypothesis

Non-Directional Relational Hypothesis

Directional Relational Hypothesis

Directional Causal Hypothesis

Stronger

HYPOTHESES EXAMPLE

- The query response time is measured for a few different search engines
- Different hypotheses
  - **Descriptive:** The query response time for Google follows a normal distribution
  - **Non-directional relational:** The average response time for a new search engine, QuickSearch, is different from Google's average response time
  - **Directional relational:** The average response time of QuickSearch is shorter than that of Google's
  - **Directional causal:** The response time of QuickSearch is shorter than Google's because they cache results of more queries

REMINDER & NEXT CLASS

- Reminder: Assignment 1 is due on Sunday (Jan 20), 11:59pm
  - You can not apply extension days on Assignment 1!
- Next class: Elements of data mining algorithm