Computer Science

CSC380: Principles of Data Science

Statistics 1

Xinchen Yu

- Probability
- Statistics

- Data Visualization
- Predictive modeling
- Clustering

Outline

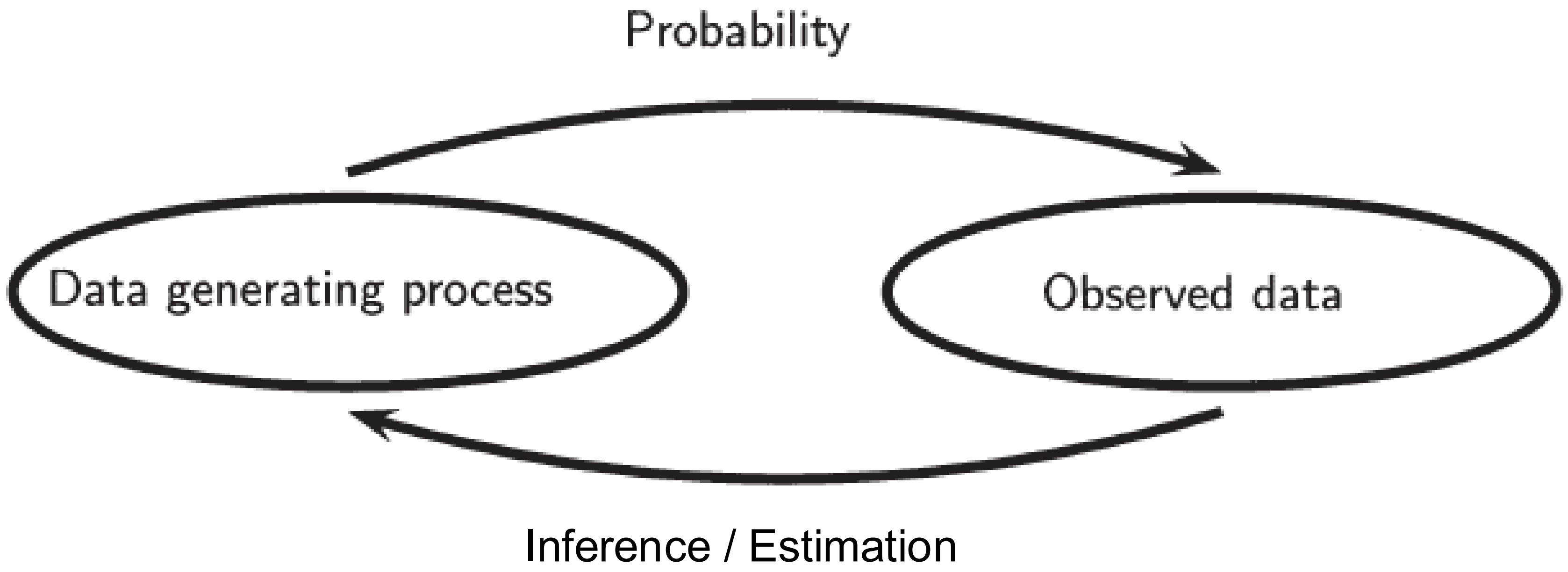
- Basic setup of parameter estimation
- Plug-in estimators
- Maximum-likelihood estimators

Probability and Statistics

*Probability: **Given a distribution**, compute probabilities of data/events.*

E.g., Given 5 fair coin flips, what is the probability of #heads ≥ 3?

e.g., data = outcome of coin flip



E.g., We observed 5 flips of a coin H, T, T, T, T. How fair is the coin?

*Statistics: **Given data**, compute/infer the distribution or its properties.*

[Source: Wasserman, L. 2004]

Intuition Check

Suppose that we toss a coin 100 times. We don't know if the coin is fair or biased…

Question 1 Suppose that we observe **52** heads and **48** tails. Is the coin fair? Why or why not?

Perhaps fair

Question 2 Now suppose that out of 100 tosses we observed **73** heads and **27** tails. Is the coin fair? Why or why not?

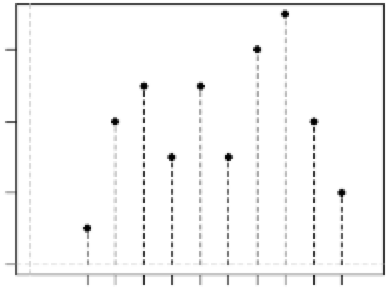
Perhaps unfair

Question 3 How to estimate the bias of the coin with **73** heads and **27** tails if using 73/100?

Let's see..

Good & bad estimators

Example Estimate $\theta = \mu = \sum_x x \cdot f(x)$ for an unknown distribution



Say true $\theta = 3.5$

Our dataset X_1, X_2, X_3, X_4 are 3,6,5,-2.

Can try to estimate θ using *any function* of $X_1, \ldots, X_4$:

$\hat{\theta}_N$:	$\frac{1}{4}\sum_{i=1}^{4} X_i$	$\frac{\min(X_1,..,X_4) + \max(X_1,..,X_4)}{2}$	$X_1 \cdot X_4$
	3	2	-6

Good & bad estimators

Given an already-drawn sample, the **quality** of an estimator depends on the *representativeness* of the sample.

e.g.

$$\frac{1}{4}\sum_{i=1}^{4} X_i \text{ or } X_1 \cdot X_4$$

Example Coin toss $X \sim \text{Bernoulli}(p = 0.5)$

- If unlucky to observe 1, 1, 1, 1, then both estimators perform badly
- When we say "$\frac{1}{4}\sum_{i=1}^{4} X_i$ is a better estimator than $X_1 \cdot X_4$", what exactly do we mean?

Estimating Coin Bias

We can model each coin toss as a Bernoulli random variable,

$X \sim \text{Bernoulli}(p)$ => PMF

x=0	x=1
1-p	p

Recall that p is the coin bias (probability of heads) and that,

$$\mathbf{E}[X] = p$$

Suppose we observe N coin flips $x_1, \ldots, x_N$, estimate p using *sample mean*

$$\hat{p} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Why is this a good guess?

Parameter Estimation: basic framework

We pose a <u>model</u> in the form of a probability distribution, with unknown **parameters of interest $\boldsymbol{\theta}$,**

$$p_\theta$$

e.g. biased coin:
$\theta = p$
p_θ: Bernoulli(p)

Observe a sample of N *independent identically distributed (iid)* data points

$$x_1, \dots, x_N \sim p_\theta,$$

e.g. first sample: 1, 0, 0, 0, 0
second sample: 0, 1, 0, 1, 1

Find an **estimator** to estimate parameters of interest,

$$\hat{\theta}_N = r(x_1, \dots, x_N)$$

e.g. sample mean

1/5 for the first dataset

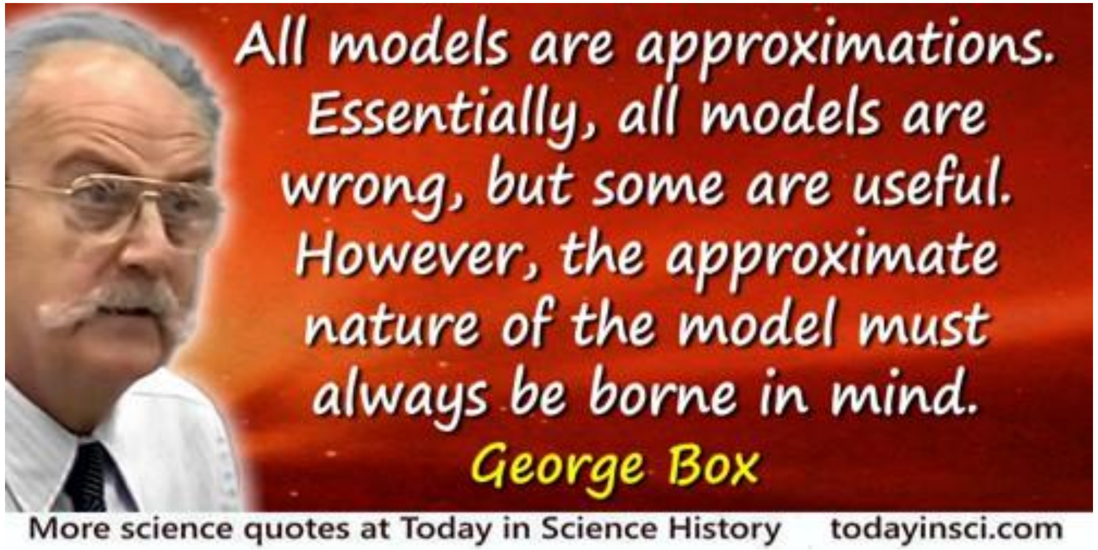
3/5 for the second dataset

Note: θ fixed and unknown; $\hat{\theta}_N$ is a random variable

Parameter Estimation: basic framework

- We pose a <u>model</u> in the form of a probability distribution p_θ, with unknown **parameters of interest** $\boldsymbol{\theta}$

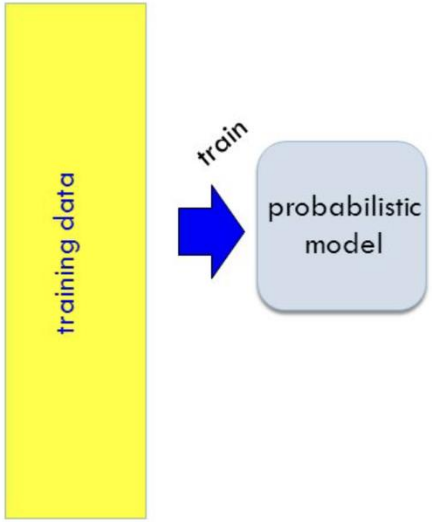
- Where do such models come from?
- Models are found by trial and errors in different applications



Bigger picture: Connection to Machine Learning

Statistical inference is sometimes called "probabilistic machine learning":

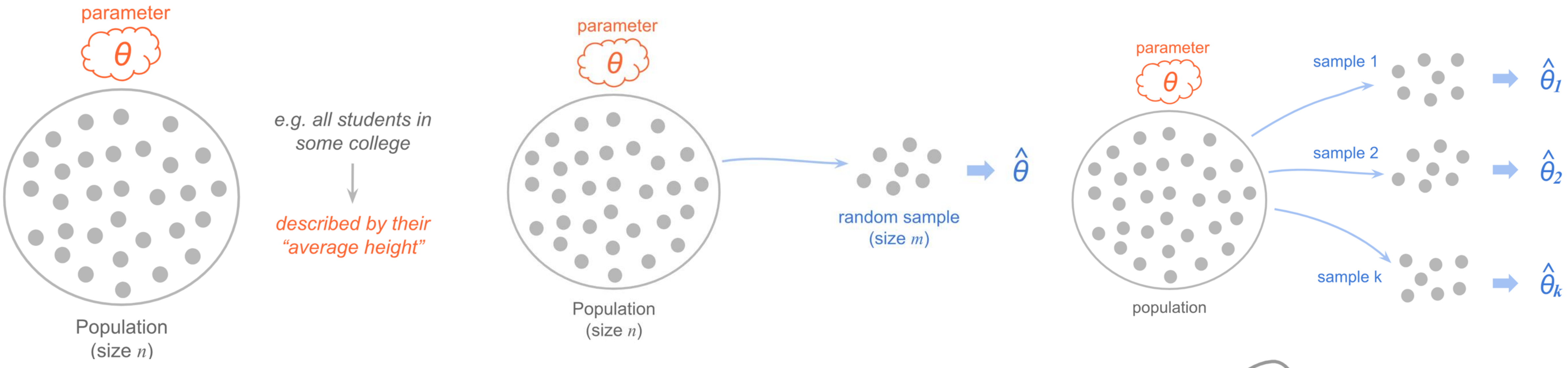
1. Model how the data is generated by probabilistic models, but with parameters unspecified (modeling assumption / generative story)
2. (Training) Learn the model parameter $\hat{\theta}$
3. (Test) Make prediction / decision based on the learned model $P(z; \hat{\theta})$



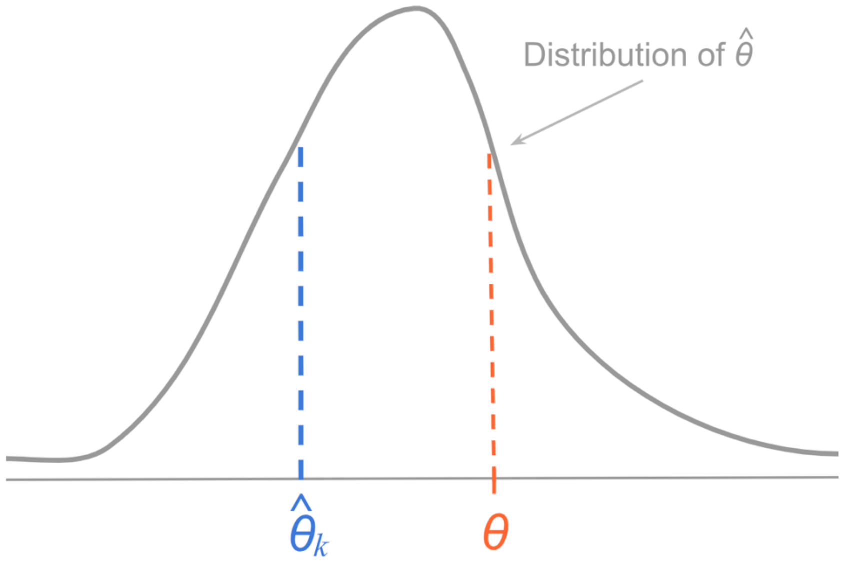
In Statistics, we mostly stop at step 2

Machine Learning cares more about step 3: prediction & decision

Distribution of an estimator



$\hat{\theta}_n$ *is a random variable, it has a distribution*

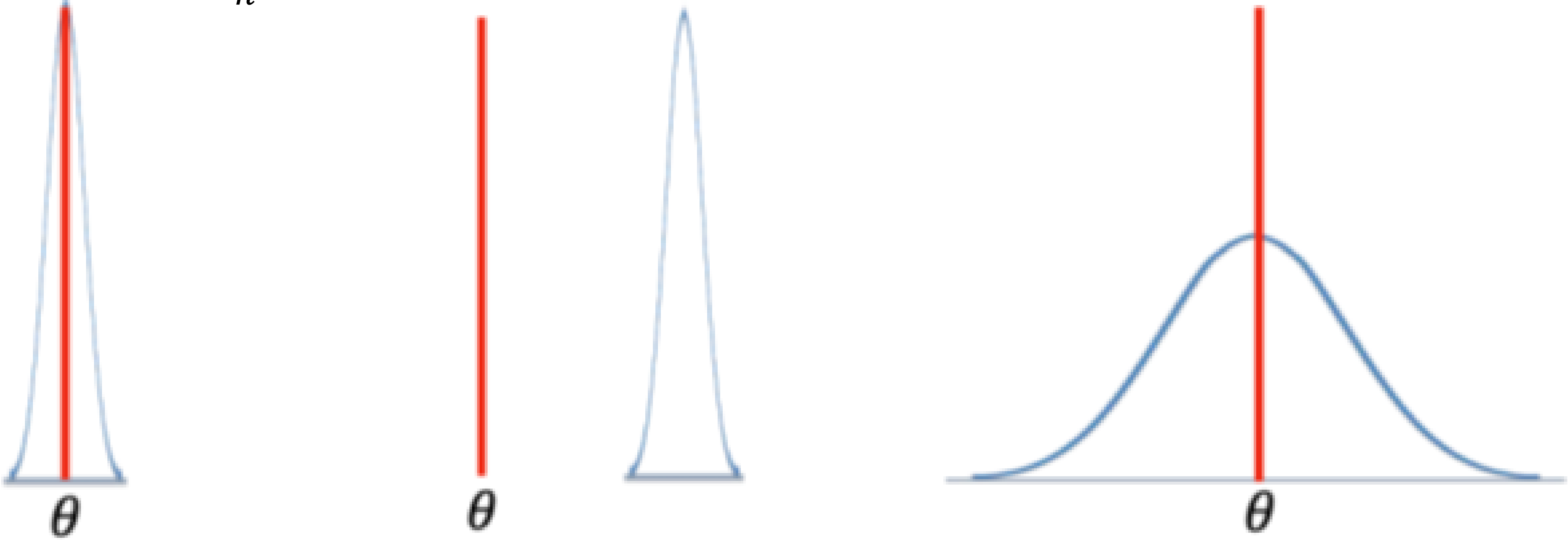


How good is an estimator

- We can get a sense of the quality of an estimator $\hat{\theta}_n$ by plotting its *probability distribution*

Recall: $\hat{\theta}_n$ *is a random variable*

Distribution of $\hat{\theta}_n$



Good

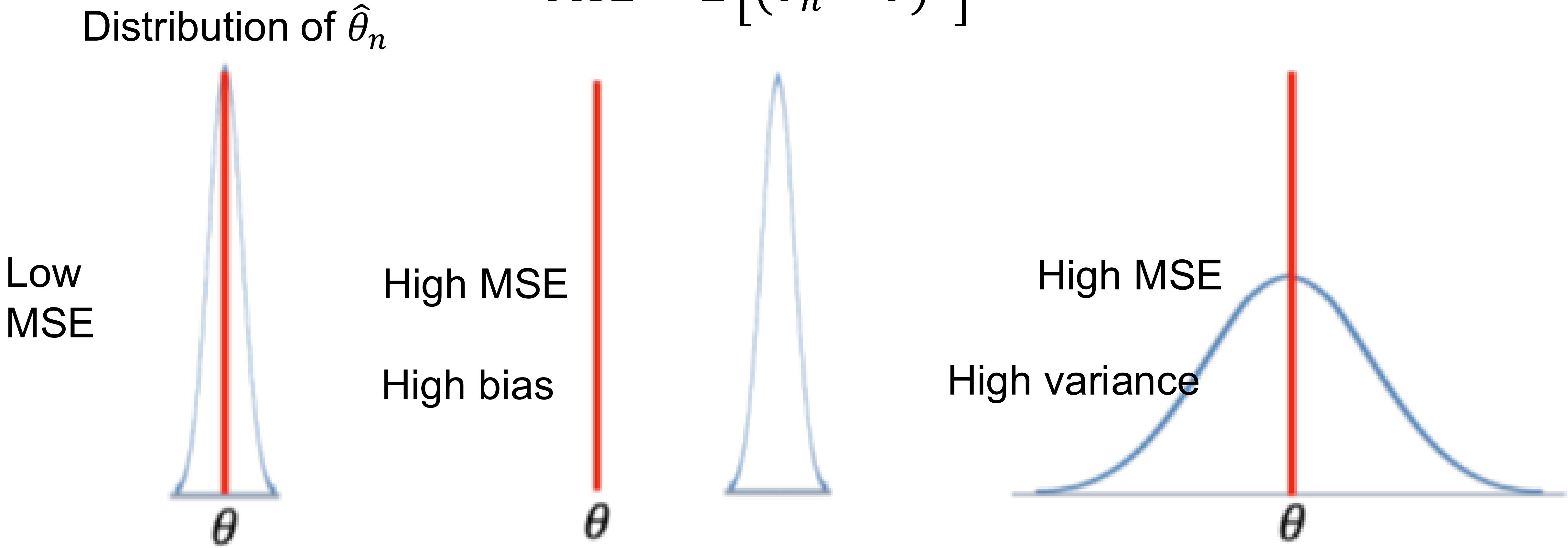
Bad

Bad

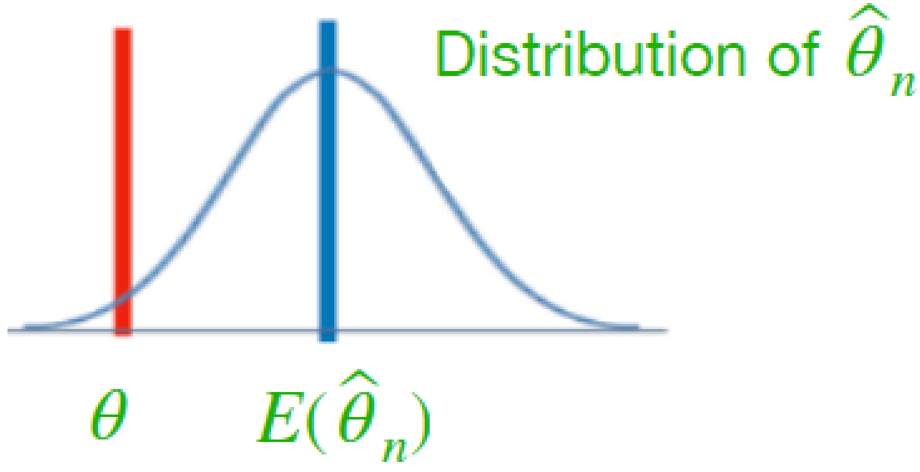
How good is an estimator

- Quantitatively, we can use the mean squared error (MSE) to measure the quality of an estimator

$$\text{MSE} = \text{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right]$$



Bias



- Bias: expected overestimate of θ

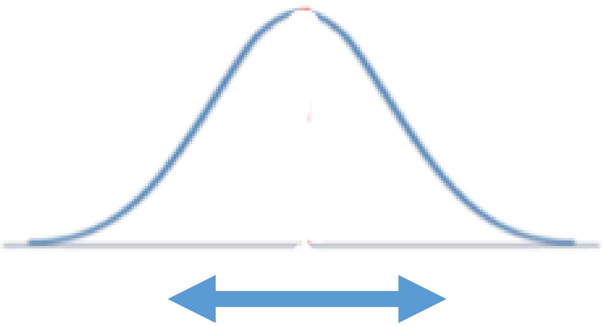
- $\mathrm{Bias}(\hat{\theta}_n) = \mathrm{E}[\hat{\theta}_n] - \theta$

  also denoted as $\mu_{\hat{\theta}_n}$

- An estimator is *unbiased* if $\mathrm{Bias}(\hat{\theta}_n) = 0$

Variance

Distribution of $\hat{\theta}_n$



- Variance: how much $\hat{\theta}_n$ deviate from its mean

- $\mathrm{Var}\left(\hat{\theta}_n\right) = \mathrm{E}\left[(\hat{\theta}_n - \mathrm{E}\left[\hat{\theta}_n\right])^2\right]$

The bias-variance decomposition

Fact The MSE of an estimator $\hat{\theta}_n$ can be decomposed as:

$$\mathrm{MSE} = \mathrm{Bias}\left(\hat{\theta}_n\right)^2 + \mathrm{Var}\left(\hat{\theta}_n\right)$$

Justification

$\mu_{\hat{\theta}_n}$: the mean of $\hat{\theta}_n$

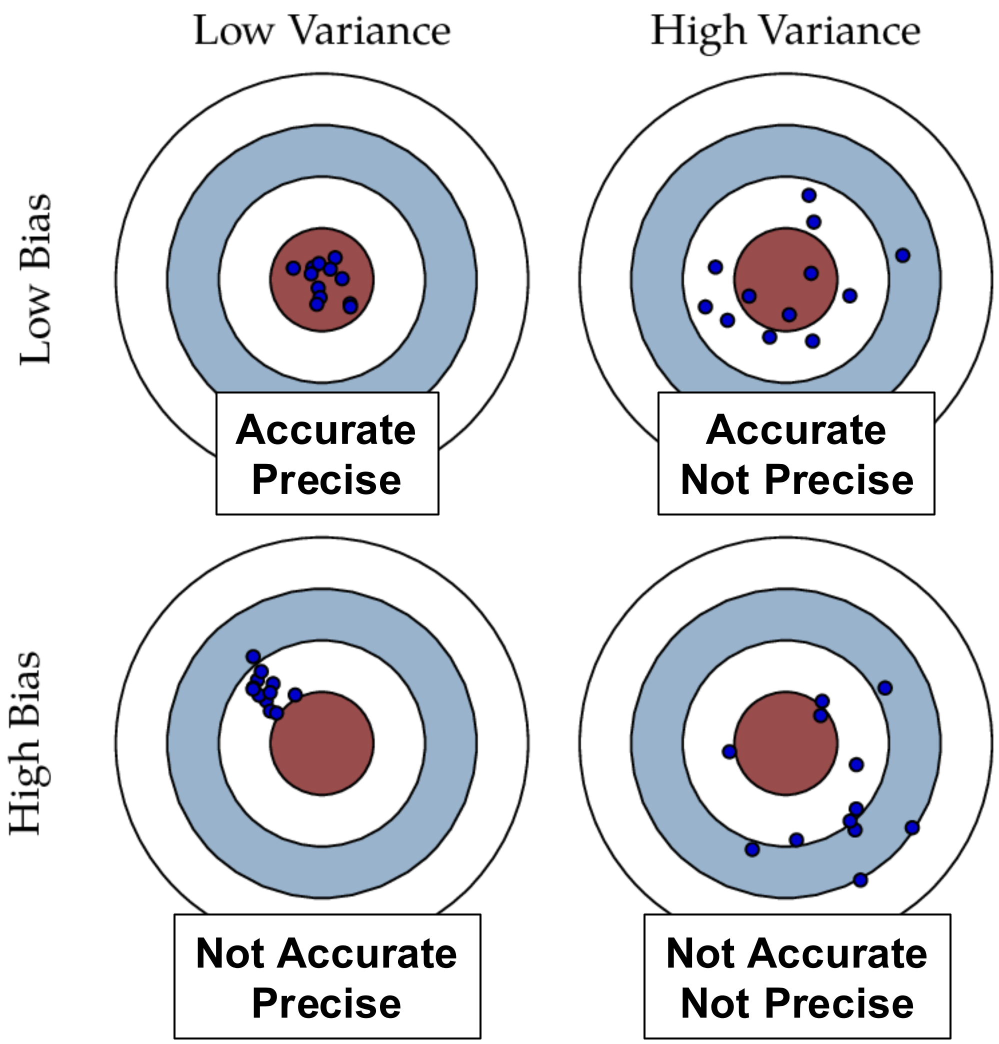
$$\mathrm{MSE} = \mathrm{E}\left[(\hat{\theta}_n - \mu_{\hat{\theta}_n} + \mu_{\hat{\theta}_n} - \theta)^2\right]$$

$$= \mathrm{E}\left[(\hat{\theta}_n - \mu_{\hat{\theta}_n})^2 + (\mu_{\hat{\theta}_n} - \theta)^2 + 2(\hat{\theta}_n - \mu_{\hat{\theta}_n})(\mu_{\hat{\theta}_n} - \theta)\right]$$

Variance — Bias — 0 (why?)

Bias and Variance

Suppose an archer takes multiple shots at a target…



$$\text{MSE} = \text{Bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)$$

- Target = θ
- Each shot = an estimate $\hat{\theta}$

- Bias ≈ systematic error
- Variance ≈ random error

Coinflip

Example Observe n coin flips $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

We use the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate p. Find this estimator's bias, variance, MSE.

$$\text{E}[X_i] = p$$
$$\text{Var}[X_i] = p(1-p)$$

$$\text{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \text{E}[X_i] = p \Rightarrow \text{Bias} = 0$$

$$\text{Var}[\bar{X}_n] = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}[X_i] = \frac{p(1-p)}{n}$$

$$\text{MSE} = \text{Bias}^2 + \text{Variance} = \frac{p(1-p)}{n}$$

Coinflip: Laplace's estimator

Example Observe n coin flips $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

Consider another estimator $\hat{p}_B = \frac{1+\sum_i X_i}{2+n}$

e.g. 7 successes out of 10 trials,

sample mean $\bar{X}_n$: $\frac{7}{10} = 0.7$

new estimator $\hat{p}_B$: $\frac{8}{12} = 0.67$

This is called "Laplace's Law of Succession" estimator

Laplace (1814) used it to estimate the probability of sun rising tomorrow

In-class exercise: bias & variance of Laplace's estimator

Example Observe n coin flips $X_1, \dots, X_n \sim \text{Bernoulli}(p)$

Consider another estimator $\hat{p}_B = \frac{1+\sum_i X_i}{2+n}$.

Find the bias and variance of $\hat{p}_B$.

Solution

$$\mathrm{E}[\hat{p}_B] = \frac{1+\mathrm{E}[\sum_i X_i]}{2+n} = \frac{1+np}{2+n} \Rightarrow \text{Bias} = \frac{1-2p}{2+n}$$

A biased estimator

$$\text{Var}[\hat{p}_B] = \text{Var}\left[\frac{\sum_i X_i}{2+n}\right] = \frac{1}{(2+n)^2}\sum_{i=1}^{n} \text{Var}[X_i] = \frac{n\,p(1-p)}{(2+n)^2}$$

Smaller than that of sample mean: $\frac{p(1-p)}{n}$

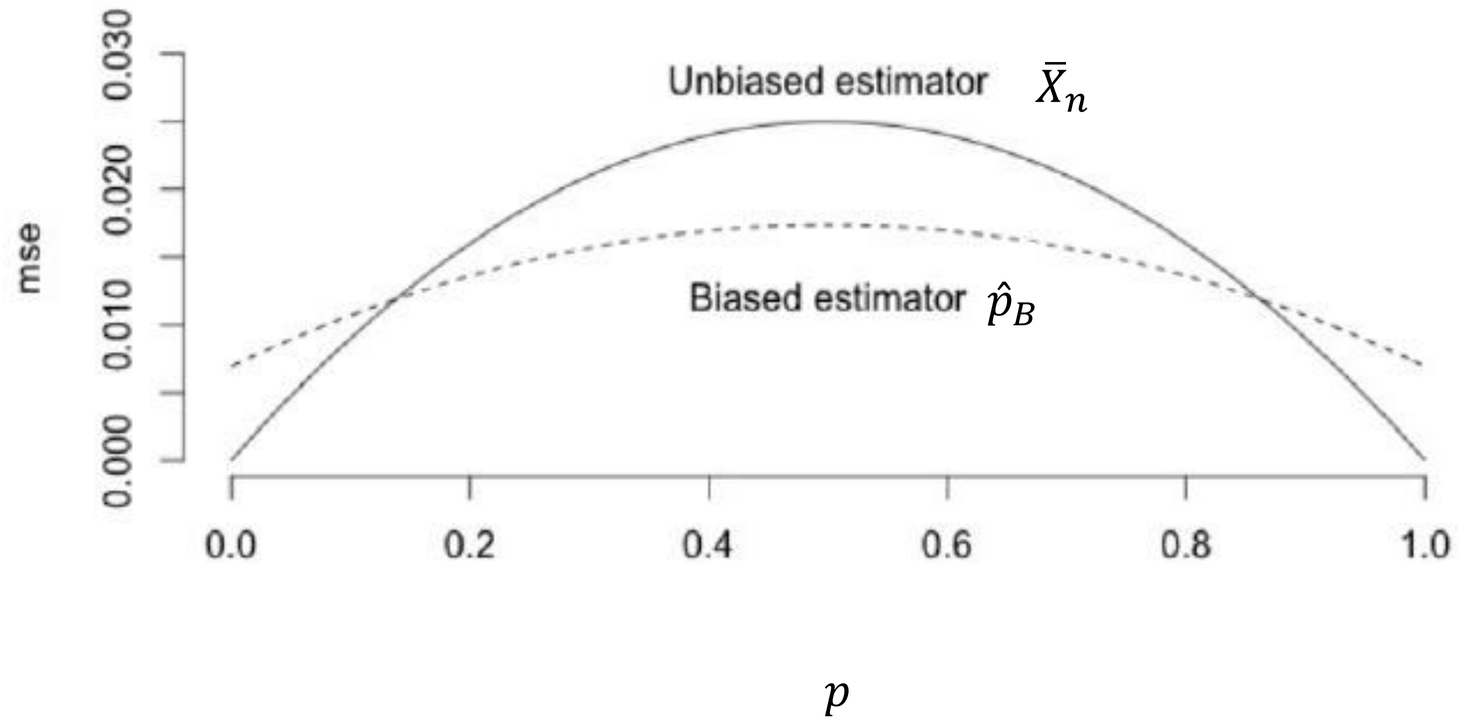
MSE = $\text{Bias}^2 + \text{Variance} = \cdots$

Comparison of two estimators

- Let's compare the two MSEs with n=10

- MSE of $\bar{X}_n$: $\frac{p(1-p)}{10}$

- MSE of $\hat{p}_B$: $\frac{1+6p-6p^2}{144}$



Is an unbiased estimator "better" than a biased one? It depends…

Warmup question: coinflips

Example Observe n coin flips $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

Consider a "blind" estimator $\hat{p} = \frac{1}{2}$.

What is $\hat{p}$'s bias and variance?

$$\text{Bias}(\hat{p}) = \text{E}[\hat{p}] - p = \frac{1}{2} - p$$

$$\text{Variance}(\hat{p}) = 0$$

$$\text{MSE}(\hat{p}) = \text{Bias}(\hat{p})^2 + \text{Variance}(\hat{p}) = \left(\frac{1}{2} - p\right)^2$$