

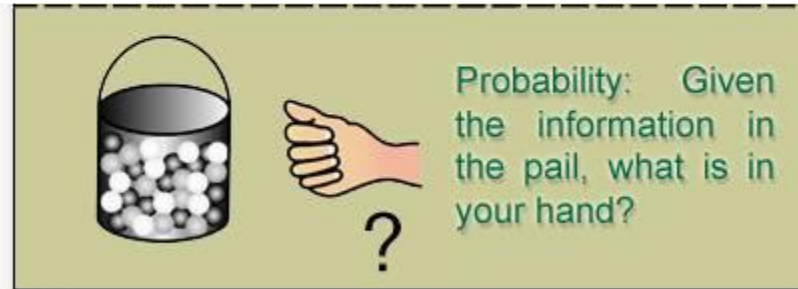


Computer
Science

CSC380: Principles of Data Science

Statistics 1

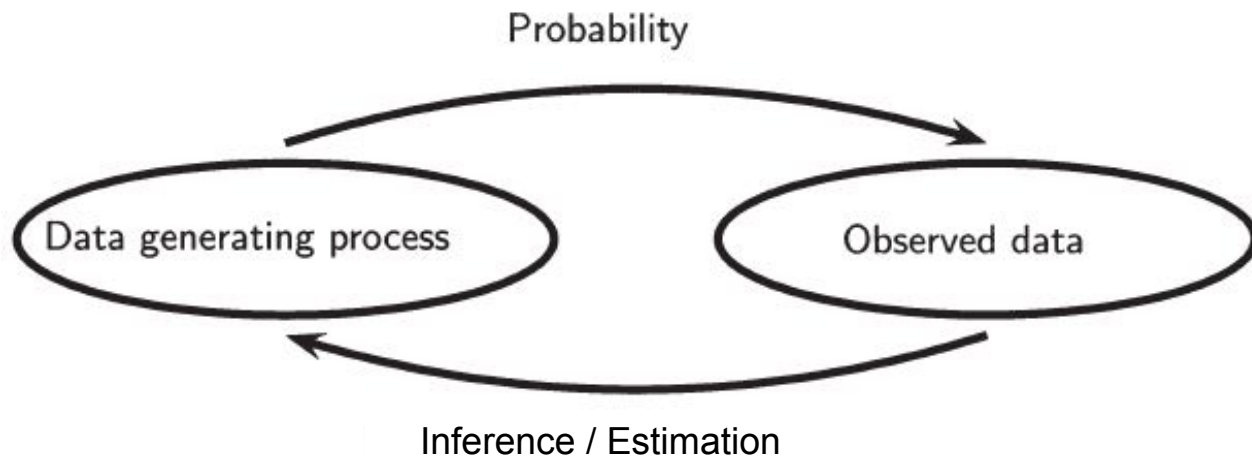
Xinchen Yu



- Probability provides a mathematical formalism to reason about **random events**
 - Knowing the distribution (e.g., uniform), how can we compute probability of the event of interest? (e.g., two fair dice, $P(\text{sum} = 3 \mid X_1 = 1)$)
- Statistics is centered on **data**
 - Fitting models to data (estimation)
 - **E.g.**, I don't know the distribution, but I have samples drawn from it. Let's estimate what the distribution was! \Rightarrow **reverse engineering!**
 - Answering questions from data (statistical inference, hypothesis testing)
 - Interpretation of data
- Statistics *uses* probability to address these tasks

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

E.g., If $X_1, \dots, X_{10} \sim \text{Bernoulli}(p=.1)$, what is the probability of $\sum_{i=1}^{10} X_i = 3$? e.g., data = outcome of coin flip



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

E.g., We observed $X_1 = 0, \dots, X_{10} = 1$. What is the head probability?

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

Question 1 Suppose that out of 100 tosses we observed 73 heads and 27 tails. What is the coin bias?

Question 2 How might we estimate the bias of the coin with 73 heads and 27 tails?



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

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

$$\mathbf{E}[X] = \pi \cdot 1 + (1 - \pi) \cdot 0 = \pi$$

Suppose we observe N coin flips x_1, \dots, x_N , estimate π as,

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

e.g. $X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 1$

$$\begin{aligned}\hat{\pi} &= \frac{1}{4}(1 + 1 + 0 + 1) \\ &= \frac{3}{4} \times 1 + \frac{1}{4} \times 0 = \frac{3}{4}\end{aligned}$$

This is called empirical mean or sample mean

Estimating Gaussian Parameters

7

Suppose we observe the heights of N students at UA, and we model them as Gaussian:

$$\{x_i\}_i^N \sim \mathcal{N}(\mu, \sigma^2)$$

(Property of Gaussian: $E[X] = \mu_x$, $Var[X] = \sigma_x^2$)

How can we estimate μ ?

$$\mu = E[X] \approx \frac{1}{N} \sum_i x_i$$

Estimate μ using sample mean

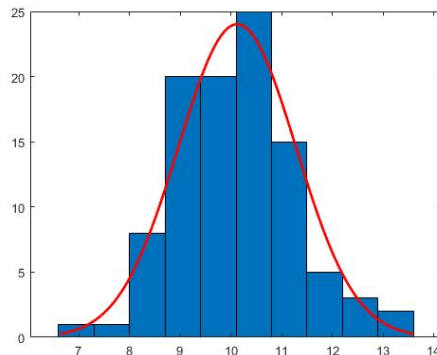
$$\hat{\mu} = \frac{1}{N} \sum_i x_i \text{ (abbrev. } \bar{x})$$

How can we estimate σ ?

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] \approx \frac{1}{N} \sum_i (x_i - \mu)^2 \approx \frac{1}{N} \sum_i (x_i - \hat{\mu})^2$$

Estimate σ using

$$\hat{\sigma} = \sqrt{\frac{1}{N} \sum_i (x_i - \hat{\mu})^2}$$



Limit Theorems: LLN and CLT

Claim: sample mean converges to the true mean.

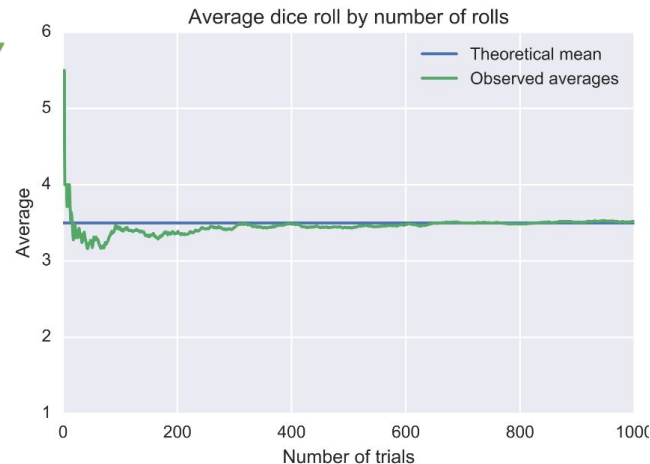
(Theorem) Let $X_1 \dots X_N$ be drawn *independent identically distributed* (i.i.d.) from a distribution with mean μ .

Then,

$$\lim_{N \rightarrow \infty} \hat{\mu}_n = \mu$$

This is the **law of large numbers**

Limitation: it does not say how does each $\hat{\mu}_n$ pile up!



Law of Large Numbers (LLN): example

10

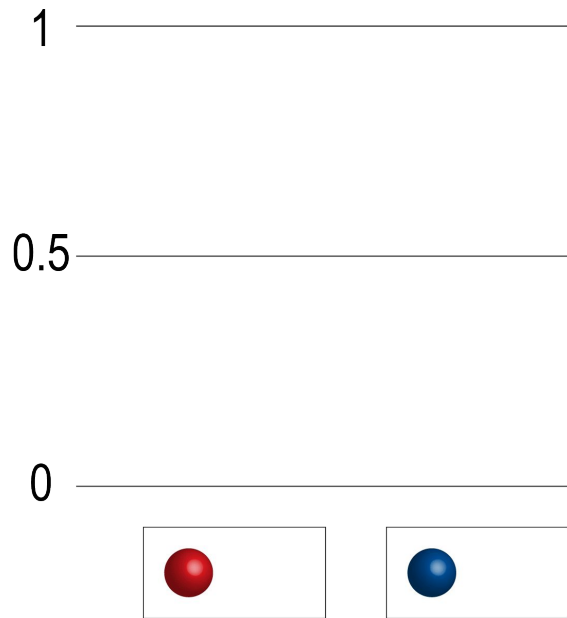
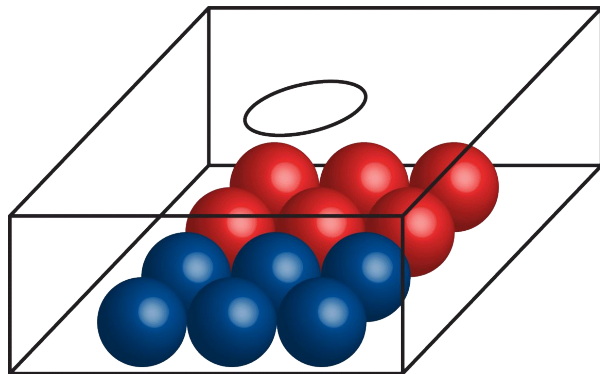
● $X = 1$

● $X = 0$

$X_1 \dots X_N \in \{0,1\}$

$\mu = 0.5$

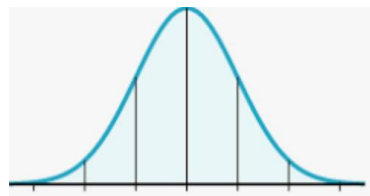
$\lim_{N \rightarrow \infty} \hat{\mu}_n = \mu = 0.5$



Probability tool: Central Limit Theorem (CLT)

11

Let X_1, \dots, X_N be i.i.d. with mean μ and variance σ^2 . Then the sample mean \bar{X}_N approaches a Normal distribution



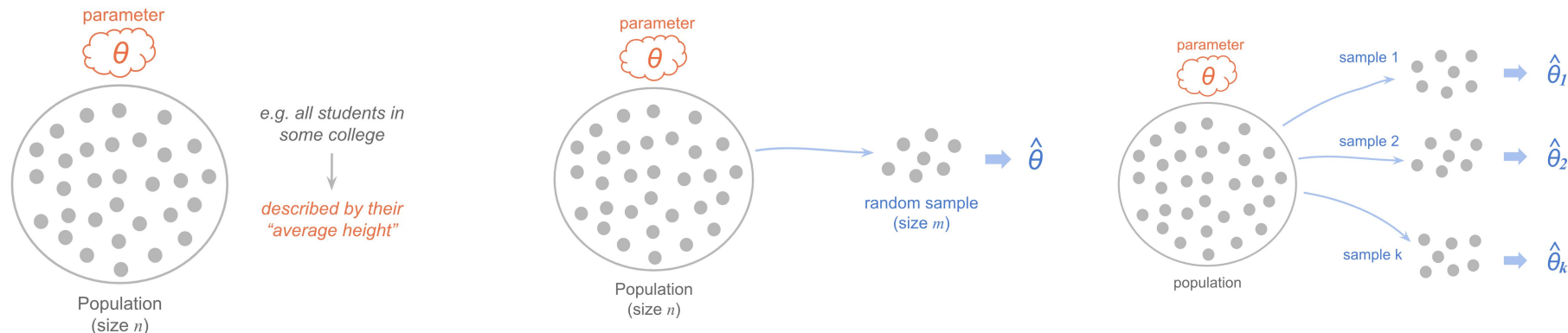
$$\lim_{N \rightarrow \infty} \bar{X}_N \rightarrow \mathcal{N} \left(\mu, \frac{\sigma^2}{N} \right)$$

Actually, a mathematically rigorous version is

$$\lim_{N \rightarrow \infty} \frac{\sqrt{N}}{\sigma} (\bar{X}_N - \mu) \rightarrow \mathcal{N}(0, 1)$$

Comments

- LLN says estimates \bar{X}_N “pile up” near true mean, CLT says *how* they pile up
- Pretty remarkable since we make **no assumption about how X_i are distributed**
- Variance of X_i **must be finite**, i.e. $\sigma^2 < \infty$

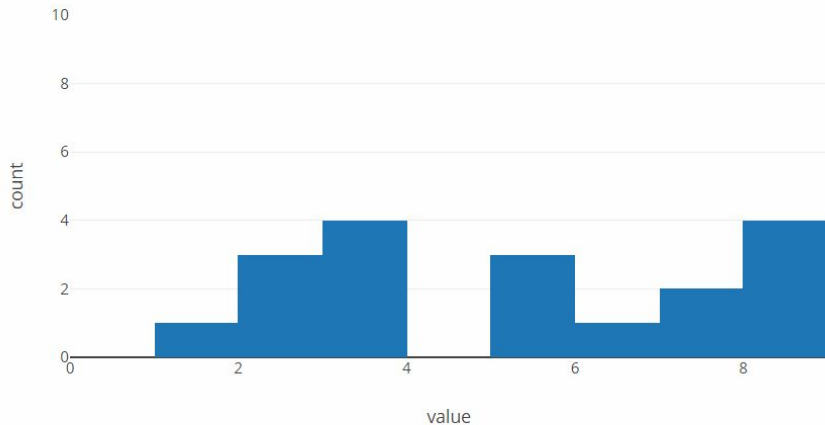


- CLT: If sampling size is large enough, the distribution of the sample mean will always be normally distributed.
- Example (next slide):
 - Sample 1: generate 20 random numbers in $[0, 9]$
 - calculate average (sample mean)
 - repeat 1000 times
 - draw distribution of the 1000 sample means

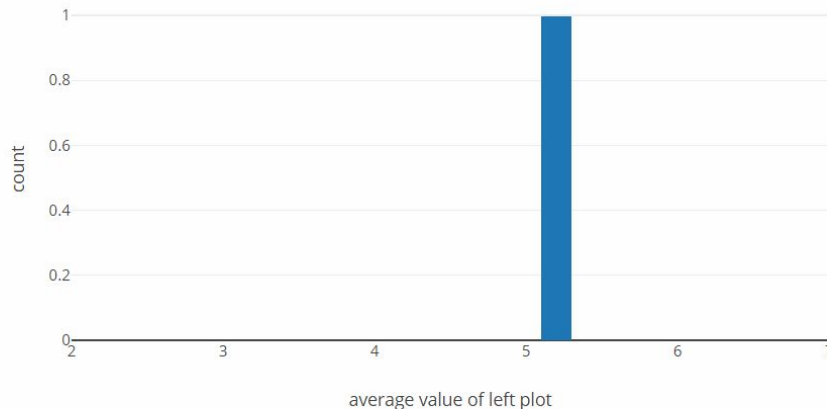
Central Limit Theorem (CLT): example

13

Random Numbers

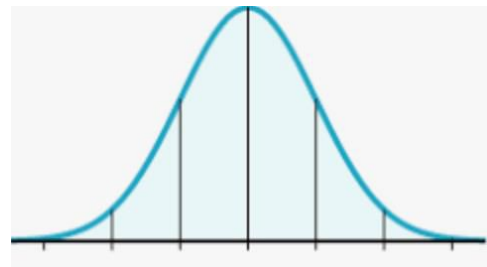


Distribution of Averages



Generate 20 random numbers from 0 and 9. Find their average. Repeat 1000 times. The averages will approximate a normal distribution (bell curve) centered at 4.5.

- Let $X_1 \dots, X_N$ be drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$
- What's the distribution of \bar{X}_N ?



$$\Rightarrow \sum_{i=1}^N X_i \sim \mathcal{N}(N\mu, N\sigma^2)$$
$$\begin{aligned} \bar{X}_N - \mu &\sim N\left(0, \frac{\sigma^2}{N}\right) \\ \frac{\sqrt{N}(\bar{X}_N - \mu)}{\sigma} &\sim N\left(\frac{\sqrt{N}}{\sigma} 0, \frac{N}{\sigma^2} \frac{\sigma^2}{N}\right) \end{aligned}$$

$$\Rightarrow \bar{X}_N \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

$$\Leftrightarrow \frac{\sqrt{N}(\bar{X}_N - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

Recall: for normal distributions

- Closed under additivity:
 $X \sim \mathcal{N}(\mu_x, \sigma_x^2) \quad Y \sim \mathcal{N}(\mu_y, \sigma_y^2) \quad , \quad X \perp Y$
 $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$
- Closed under affine transformation (a and b constant):
 $aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$

Estimation: Classical Statistics

We pose a model in the form of a probability distribution, with unknown **parameters of interest** θ ,

\mathcal{D}_θ e.g., assume Gaussian: $\theta = (\mu, \sigma^2)$

Observe data, typically *independent identically distributed (iid)*,

i.i.d.

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

$$\hat{\theta}(\{x_i\}_i^N) \approx \theta$$

Many different types of estimators, each with different properties

- Roll a die 10 times and record how many times the result is 1 (**I.I.D.**).
 - each outcome of the die roll will not affect the next one (**Independent**).
 - each roll will have the same probability as each other roll (**Identically distributed**).
- Flip two coins A and B with different weights and record how many heads (Independent but **Nonidentically distributed**).

$$P(A = H) \neq P(B = H)$$



Dependent identical distribution

- First coin (A): fair coin
- Second coin (B):
 - if $A=H$, throw an unfair coin $P(H) = \frac{1}{4}$, $P(T) = \frac{3}{4}$
 - If $A=T$, throw an unfair coin $P(H) = \frac{3}{4}$, $P(T) = \frac{1}{4}$

	B=H	B=T	
A=H	1/8	3/8	1/2
A=T	3/8	1/8	1/2
	1/2	1/2	

(joint probability table)

- $P(A=H)=P(B=H)$ but A and B are not independent (prove it!)

In general, i.i.d. is necessary to have estimators close to the true parameter

A **statistic** is a function of the data that does not depend on any unknown parameter.

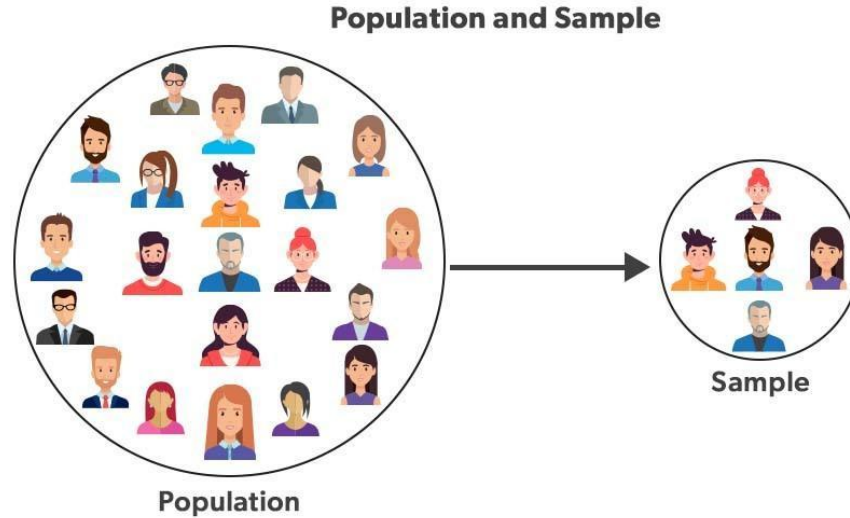
Examples

- Sample mean $\hat{\mu}$
- Sample variance $\hat{\sigma}^2$
- Sample STDEV $\hat{\sigma}$
- Order statistics $x_{(1)}, x_{(2)}, \dots, x_{(n)}$

An **estimator** $\hat{\theta}(x)$ is a **statistic** used to infer the unknown parameters of a statistical model.

Q: Gaussian distribution with unknown mean and variance.
Which of these are estimators?

A: $\hat{\mu}$ and $\hat{\sigma}$



- Parameter: mean of weight in the population μ
- Statistic: mean of weight in the sample $\hat{\mu}$
- Variable: value of weight of each person X
- Weights (kg) of people in a sample $X_1 = 80, X_2 = 60 \dots X_n = 100$
- We can use sample mean $\hat{\mu}$ to estimate population mean μ
- Sample mean is an estimator

Suppose that we toss a coin 100 times. We observe 52 heads and 48 tails...

Question 1 I define an estimator that is *always* $\hat{\theta} = 0$, regardless of the observation. Is this an estimator? Why or why not?

Question 2 Is the estimator above a **good** estimator? Why or why not?

Question 3 What are some properties that could define a **good** estimator?



- ❖ **Consistency** Given enough data, the estimator *converges* to the true parameter value

$$\lim_{n \rightarrow \infty} \hat{\theta}(x_1, \dots, x_n) \rightarrow \theta$$

Q: Is sample mean a consistent estimator for μ ?

$$\lim_{N \rightarrow \infty} \hat{\mu}_n = \mu$$

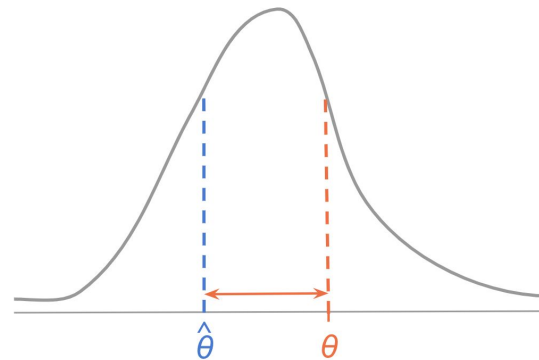
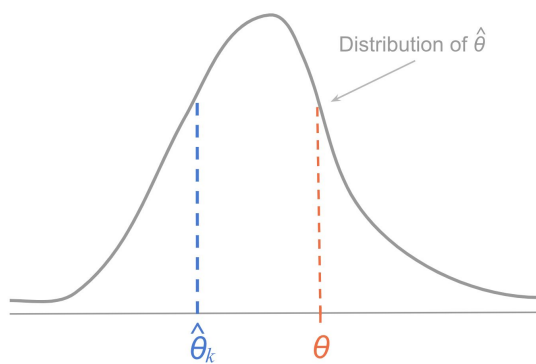
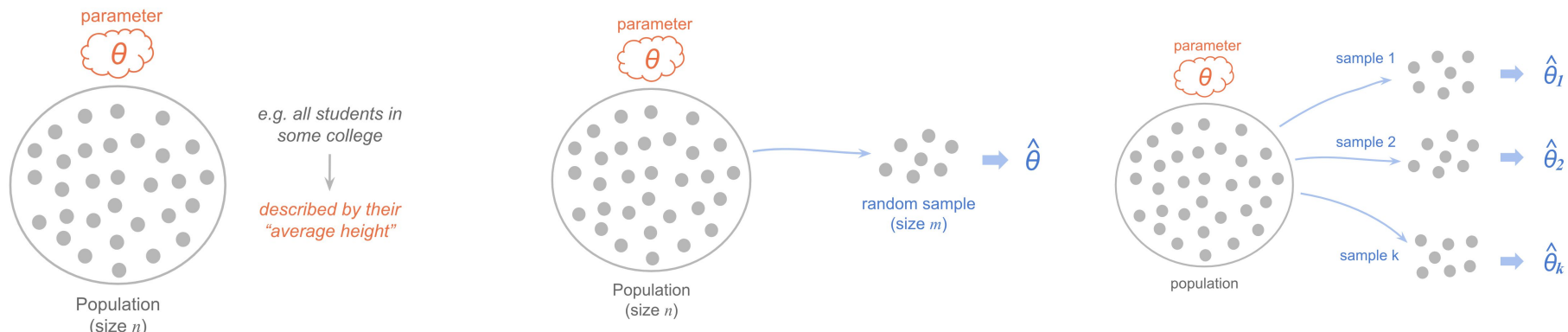
- ❖ **Efficiency** It should have low error with finite n , e.g.

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

Mean squared error should be small

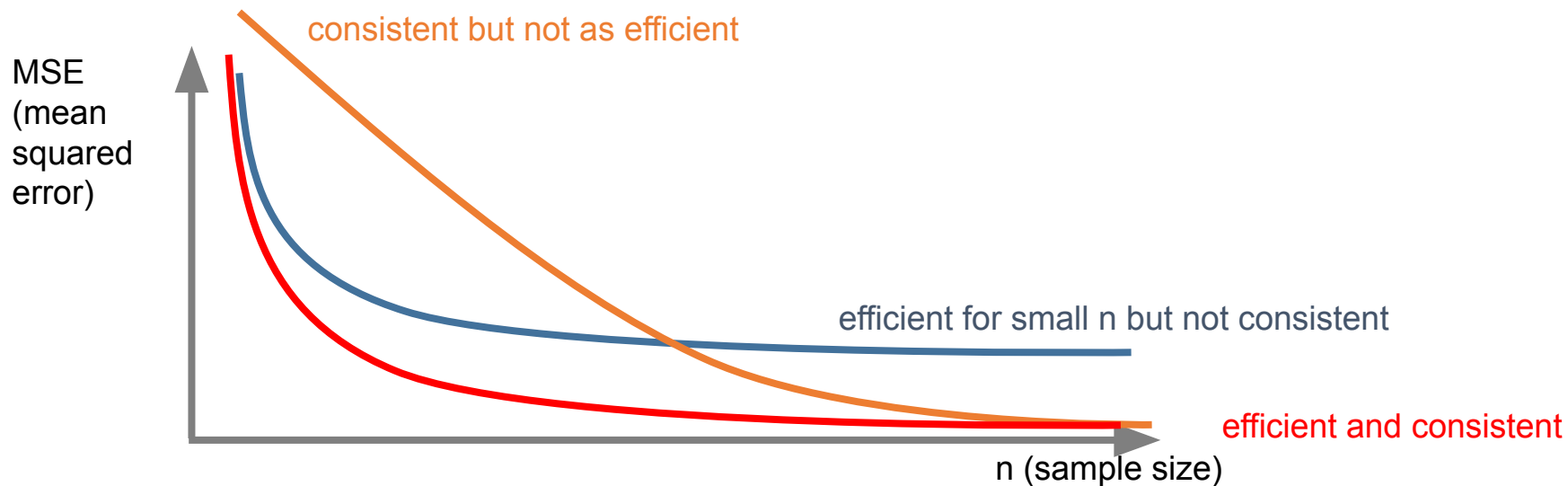
looks like variance but it's different!
Q: spot the difference from $\text{Var}(\hat{\theta}_n)$?

MSE of an Estimator



Two Desirable Estimator Properties

24



- **Unbiasedness**: For any n , $\mathbf{E}[\hat{\theta}(X_1, \dots, X_n)] = \theta$
 - E.g., sample mean is unbiased. If $X_1, \dots, X_n \sim D$ with $\mathbf{E}_{X \sim D}[X] = \mu$

$$\mathbf{E}[\bar{X}_N] = \frac{1}{N} \sum_i \mathbf{E}[X_i] = \mu$$

- Traditionally, considered to be a good property.
- In modern statistics, **not a necessary condition** to be a good estimator.
 - An unbiased estimator may be **less efficient** compared to some other **biased** estimator.
- Biased estimators can still be **consistent**.

E.g., for some estimator

$$\mathbf{E}[\hat{\theta}(X_1, \dots, X_n)] \text{ can be } \mu + \frac{1}{n}$$

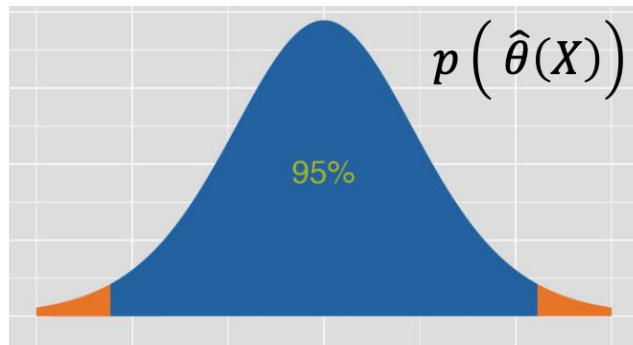
Recall: An estimator $\hat{\theta}$ is a RV (Random Variable).

Example Let $X_1, \dots, X_N \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$
and estimate \hat{p} be the *sample mean*,

$$\hat{p} = \frac{1}{N} \sum_i X_i$$

Question Is \hat{p} unbiased or not?

Notation: $X := (X_1, \dots, X_N)$



$$\mathbf{E}[\hat{p}(X)] = \mathbf{E}\left[\frac{1}{N} \sum_i X_i\right] \stackrel{(a)}{=} \frac{1}{N} \sum_i \mathbf{E}[X_i] \stackrel{(b)}{=} \frac{1}{N} Np = p$$

(a) Linearity of Expectation Operator

(b) Mean of Bernoulli RV = p

Conclusion On average $\hat{p} = p$ (it is *unbiased*)

Properties of estimators: a summary

- Consistency: when sample size N is very large, $\hat{\theta}_k$ should be very close to θ
- Efficiency: the mean squared error (expectation of squared distance between $\hat{\theta}_k$ and θ) should be small
- Unbiasedness: the expectation of $\hat{\theta}_k$ should be equal to θ

