Computer Science

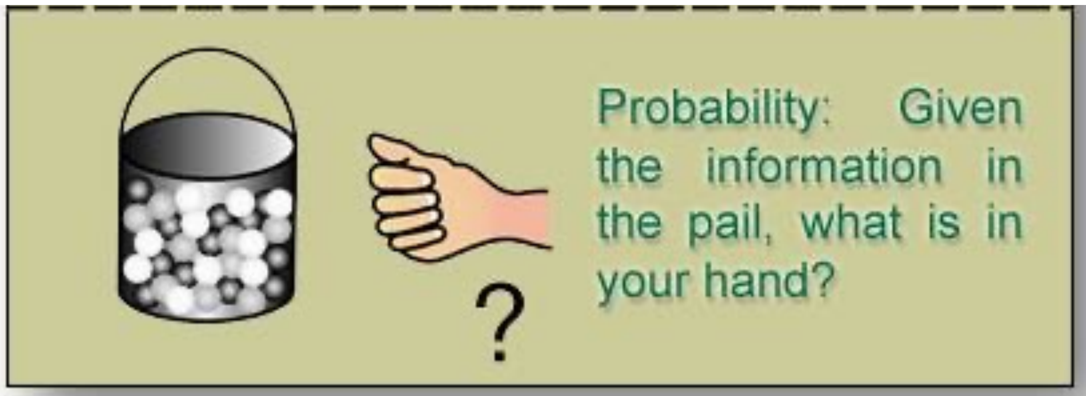
CSC380: Principles of Data Science

Statistics 1

Xinchen Yu

Acknowledgement: Built on Jason Pacheco, Kwang-Sung Jun, Chicheng Zhang's slides

Probability and Statistics



https://herdingcats.typepad.com/my_weblog/2010/04/probability-versus-statistics.html

Probability and Statistics

- 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, $P(\text{sum} = 3 \mid X_1 = 1)$)

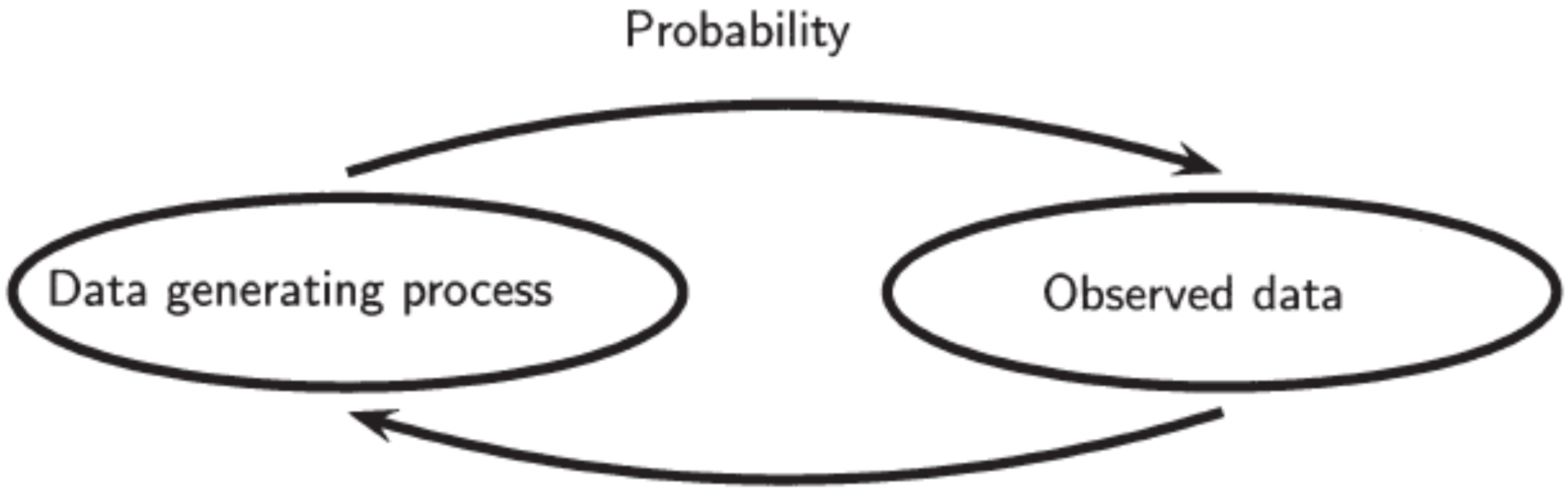
- Statistics is centered on **data**
  - Fitting models to data (estimation)
    - **E.g.,** I don't know the distribution, but I have samples drawn from it. Let's estimate what the distribution was! ⇒ **reverse engineering!**
  - Answering questions from data (statistical inference, hypothesis testing)
  - Interpretation of data

- Statistics *uses* probability to address these tasks

Probability and Statistics

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

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



Inference / Estimation

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

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

[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 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?

Estimating Coin Bias

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

$$X \sim \text{Bernoulli}(\pi) \;\; \Rightarrow \;\; p(X = x) = \pi^x (1-\pi)^{1-x}$$

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 $x_1, \ldots, x_N$, estimate π as,

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

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

$\hat{\pi} = \frac{1}{4}(1 + 0 + 0 + 1) = \frac{1}{2}$

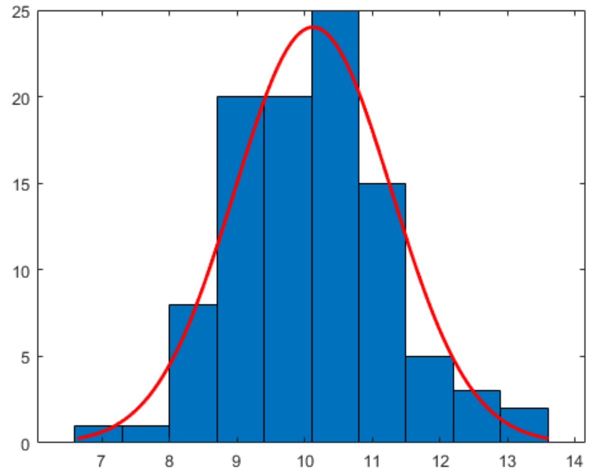
This is called <u>empirical mean</u> or <u>sample mean</u>

Estimating Gaussian Parameters

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: $\mathrm{E}[X] = \mu_x, Var[X] = \sigma_x^2$)



How can we estimate μ?

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

Estimate μ using sample mean $\hat{\mu} = \frac{1}{N}\sum_i x_i$ (abbrev. $\bar{x}$)

How can we estimate σ?

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

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

Limit Theorems: LLN and CLT

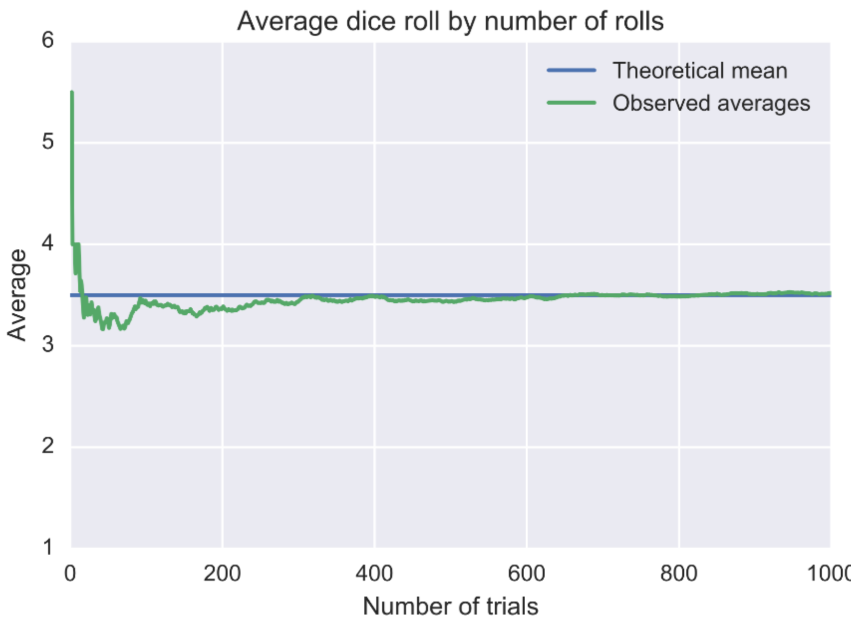
Probability tool: Law of Large Numbers (LLN)

Claim: sample mean converges to the true mean.

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

Then,

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



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

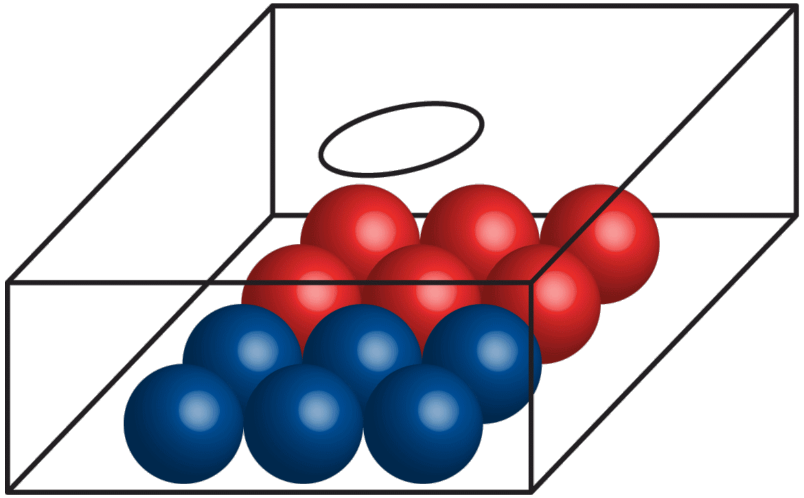
$X = 1$

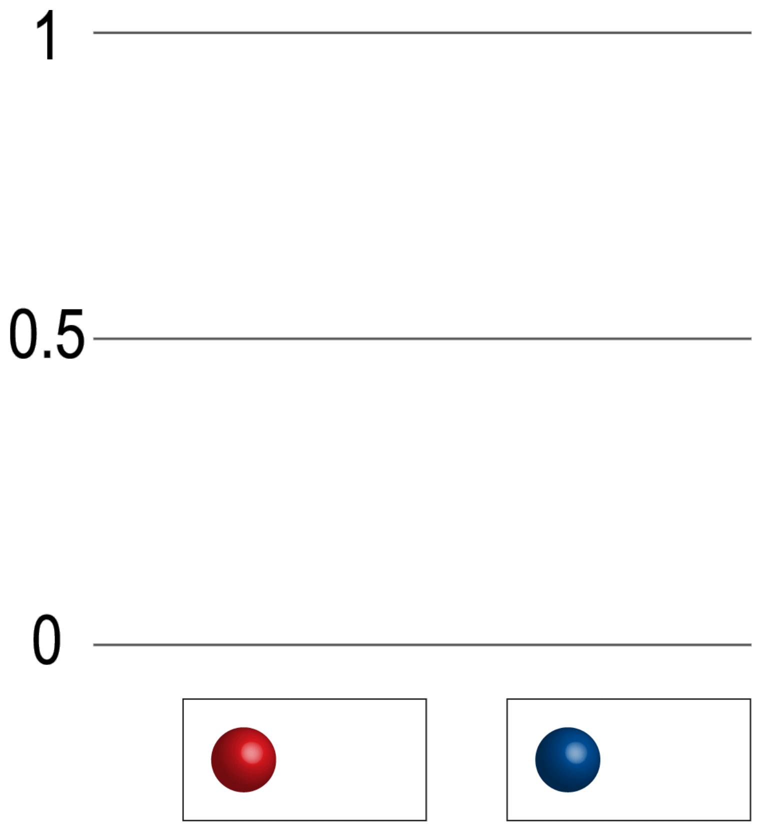
$X = 0$

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

$\mu = 0.5$

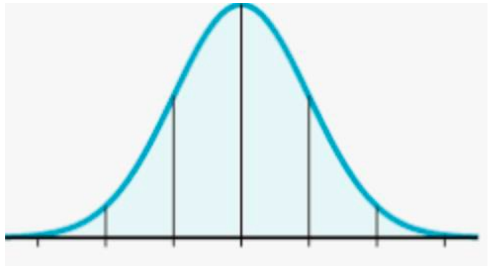
$\lim_{N \to \infty} \hat{\mu}_n = \mu = 0.5$





Probability tool: Central Limit Theorem (CLT)

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



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

=> the convergence rate is $\frac{\sigma}{\sqrt{N}}$!!

Actually, a mathematically rigorous version is

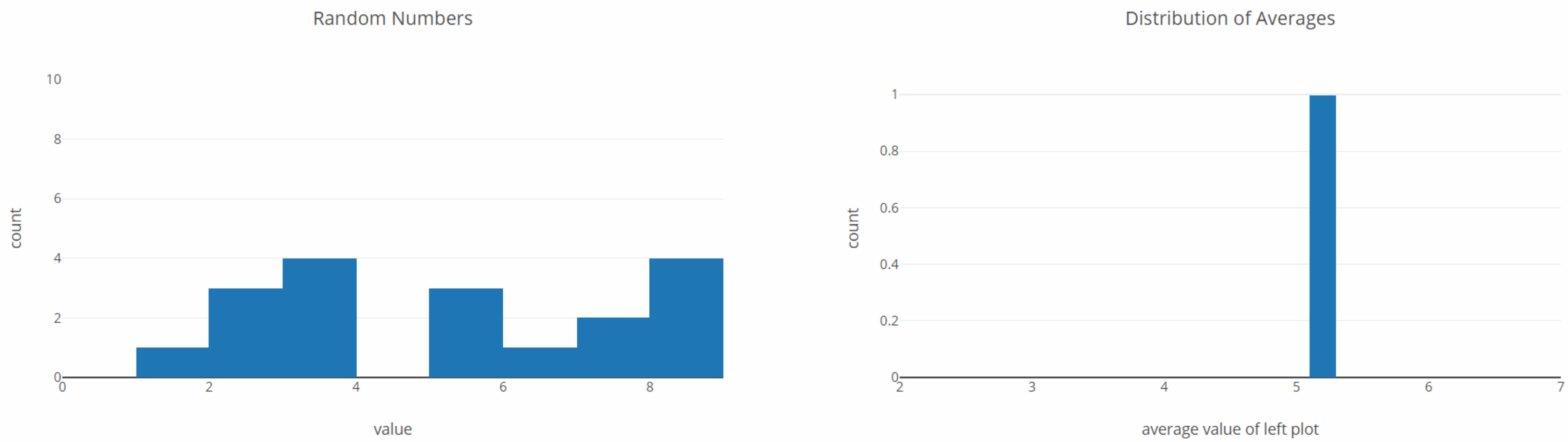
$$\lim_{N \to \infty} \frac{\sqrt{N}}{\sigma}(\bar{X}_N - \mu) \to \mathcal{N}(0, 1)$$

Comments

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

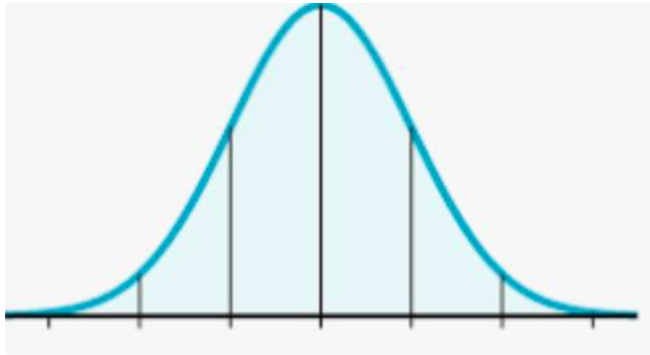
https://stats-interactives.ctl.columbia.edu/central-limit-theorem

Central Limit Theorem (CLT): example



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.

Central Limit Theorem (CLT): sanity check



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

$$\Rightarrow \sum_{i=1}^{N} X_i \sim \mathcal{N}\left(N\mu, N\sigma^2\right)$$

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

$$\overline{X_N} - \mu \sim N\left(0, \frac{\sigma^2}{N}\right)$$

$$\frac{\sqrt{N}(\overline{X_N} - \mu)}{\sigma} \sim N(\frac{\sqrt{N}}{\sigma}0, \frac{N}{\sigma^2}\frac{\sigma^2}{N})$$

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

Recall: for normal distributions

- Closed under additivity:

  $X \sim \mathcal{N}(\mu_x, \sigma_x^2) \quad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, $X \perp Y$

  $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$
- 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

Parameter Estimation

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

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

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

$p(X_1 = x_1, \dots, X_N = x_n) = p(X_1 = x_1) \cdots p(X_N = x_N)$

$$x_1, \dots, x_N \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\theta,$$

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

Examples: I.I.D. and Non-I.I.D.

- 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).

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

Examples: Non-I.I.D.

<u>Dependent</u> identical distribution

- First coin (A): fair coin
- Second coin (B):
  - if A=H, throw an **<u>unfair</u>** coin P(H) = ¼, P(T) = ¾
  - If A=T, throw an **<u>unfair</u>** coin P(H) = ¾ , P(T) = ¼

(joint probability table)

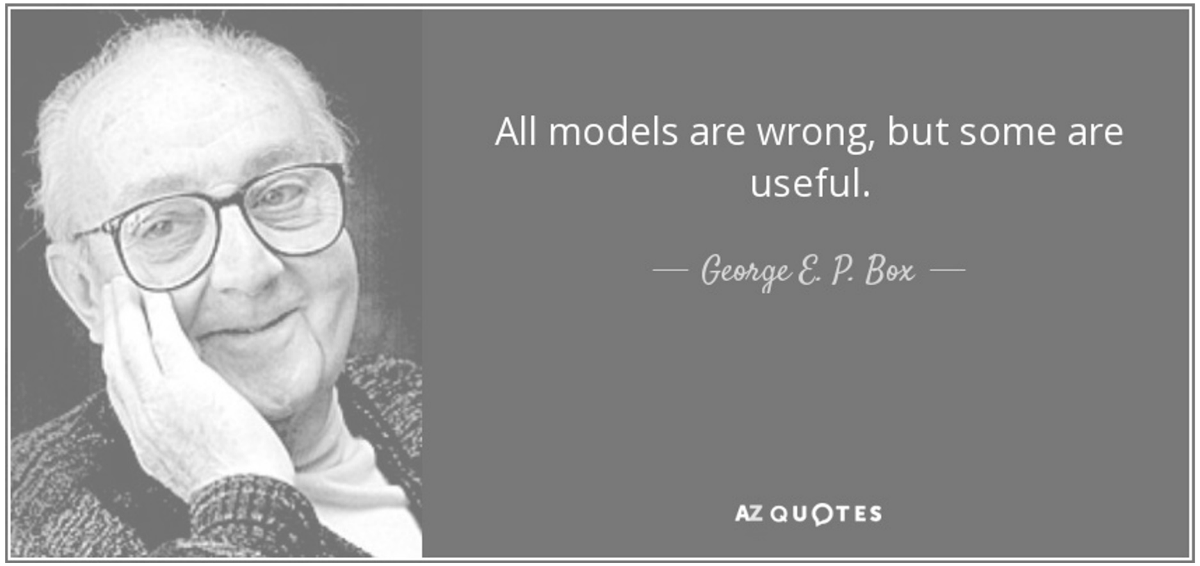
	B=H	**B=T**	
A=H	1/8	3/8	1/2
A=T	3/8	1/8	1/2
	1/2	1/2	

- 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

Models

- In the previous example, we assumed that the heights of students at UA follow a normal distribution.
- Does it?



- There are ways to check if one model is better than the other (will be covered much later)

Definitions

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

Examples

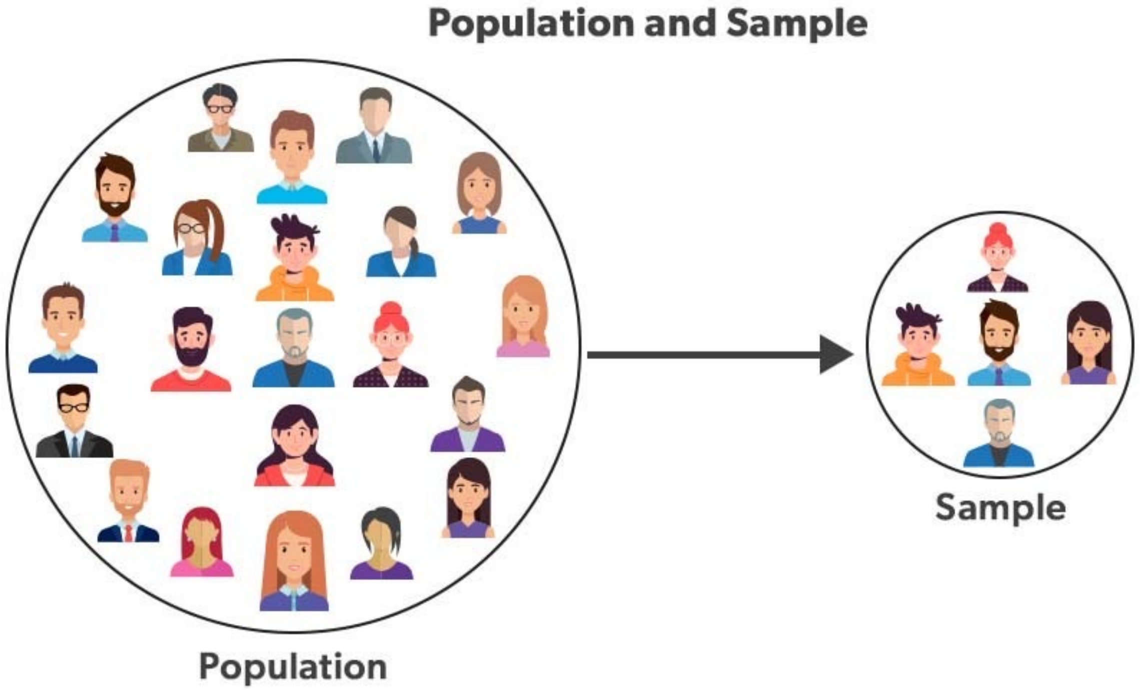
- Sample mean $\hat{\mu}$
- Sample variance $\hat{\sigma}^2$
- Sample STDEV $\hat{\sigma}$
- Order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

*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 $\hat{\sigma}$

Population and Sample



- 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 $X_1 = 80, \mathrm{X}_2 = 60 \ldots \mathrm{X}_\mathrm{n} = 100$
- We can use sample mean $\hat{\mu}$ to estimate population mean μ

Intuition Check

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?

Two Desirable Estimator Properties

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

$$\lim_{n\to\infty} \hat{\theta}(x_1, \ldots, x_n) \to \theta$$

Q: Is sample mean a consistent estimator for μ?

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

Yes. The variance of sample mean $\bar{X}_N$ decreases to 0 as we increase the sample size N.

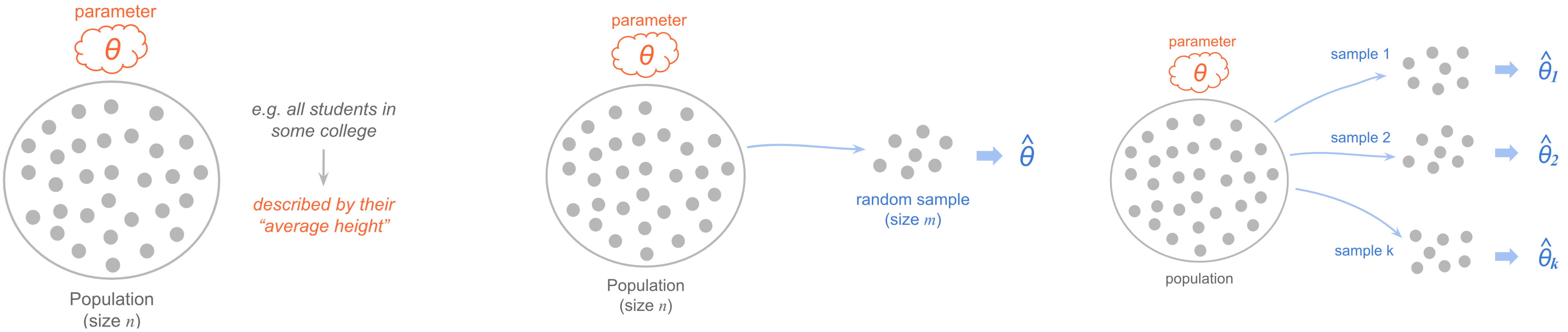
- **Efficiency** It should have low error with finite n, e.g.

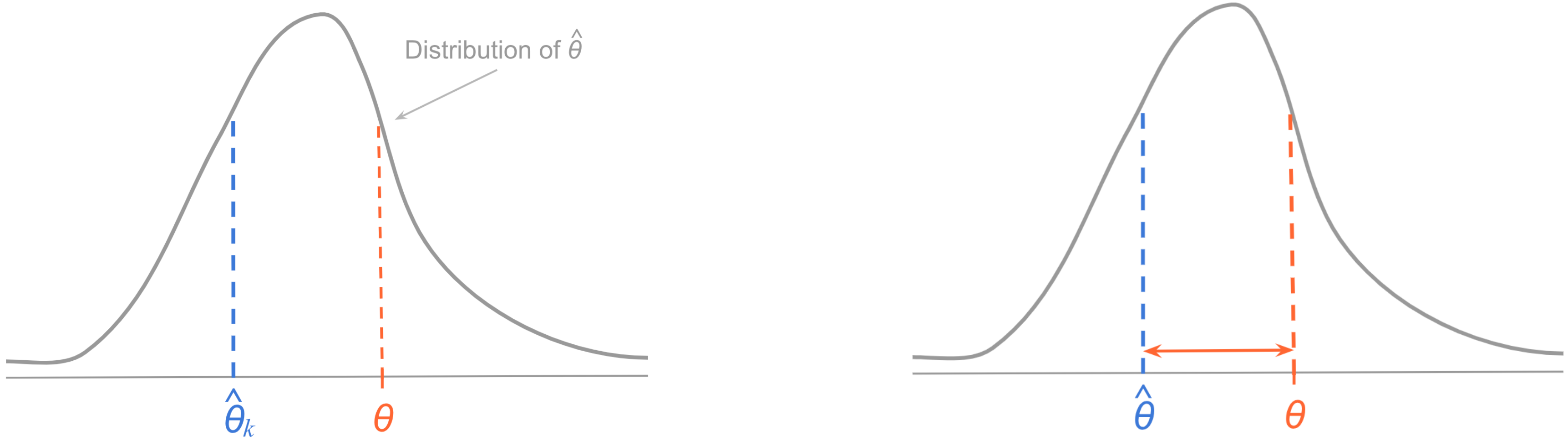
$$\mathrm{MSE}(\hat{\theta}_n) = \mathbf{E}[(\hat{\theta}_n - \theta)^2]$$

<u>Mean squared error</u> should be small

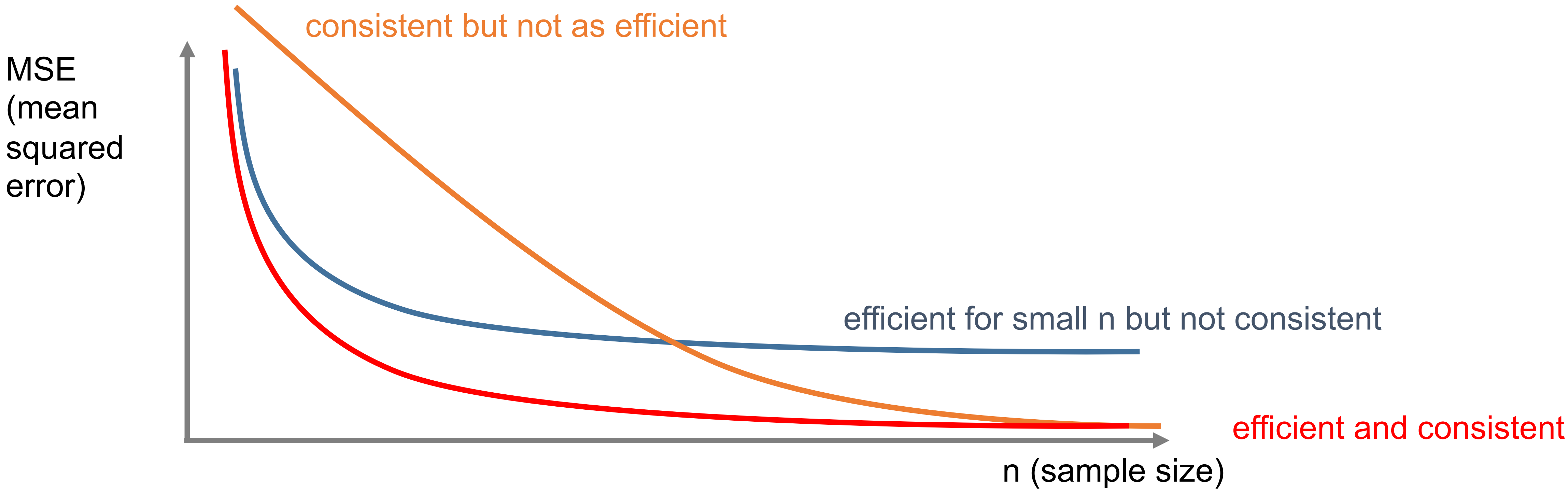
looks like variance but it's different!
Q: spot the difference from $Var(\hat{\theta}_n)$?

MSE of an Estimator





Two Desirable Estimator Properties



Another Properties of estimators

- **Unbiasedness**: For any n, $\mathbf{E}\left[\hat{\theta}(X_1, \dots, X_n)\right] = \theta$
  - 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**.

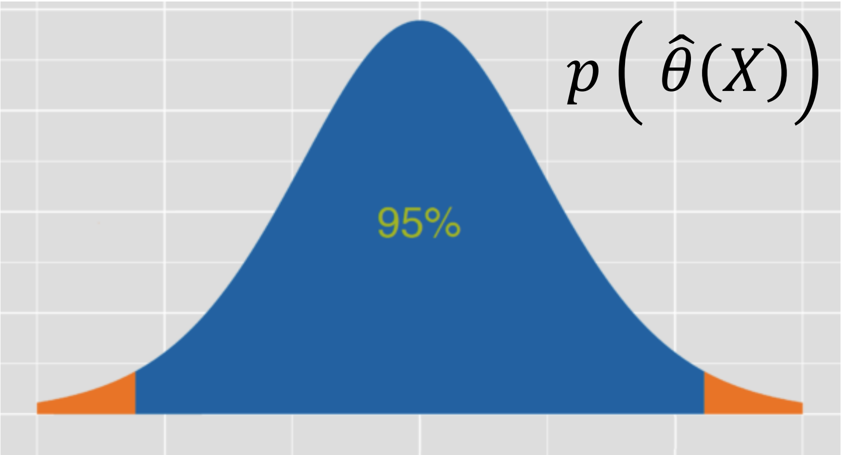
E.g., for some estimator $E\left[\hat{\theta}(X_1, \dots, X_n)\right]$ can be $\mu + \frac{1}{n}$

Expectation of the Sample Mean

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

Example Let $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$
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: $X := (X_1, \ldots, X_N)$

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

(a) Linearity of Expectation Operator

(b) Mean of Bernoulli RV = p

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