



Computer
Science

CSC380: Principles of Data Science

Statistics 1

Xinchen Yu



- Probability
- Statistics



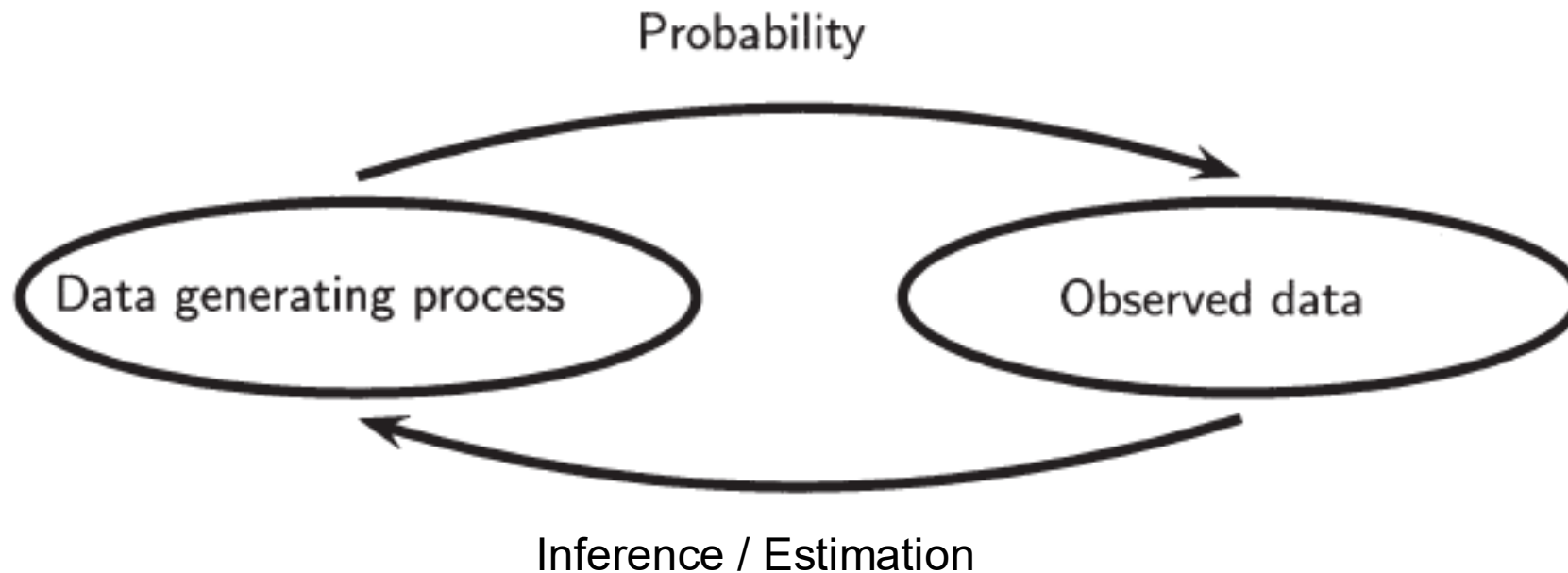
- Data Visualization
- Predictive modeling
- Clustering

- Basic setup of parameter estimation
- Plug-in estimators

*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.*

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..

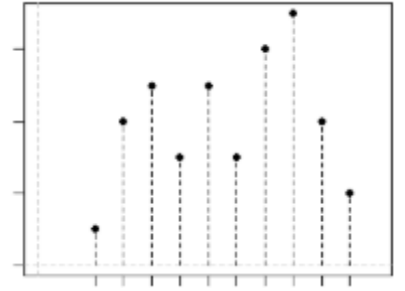


Example Estimate $\theta = \mu = \sum_x x \cdot f(x)$ for an unknown distribution

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, \dots, X_4 :



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

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

e.g.

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

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?
- How to measure the quality of an estimator?

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

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

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, \dots, x_N , estimate p using sample mean

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

Why is this a good guess?

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

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

$$p_\theta$$

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

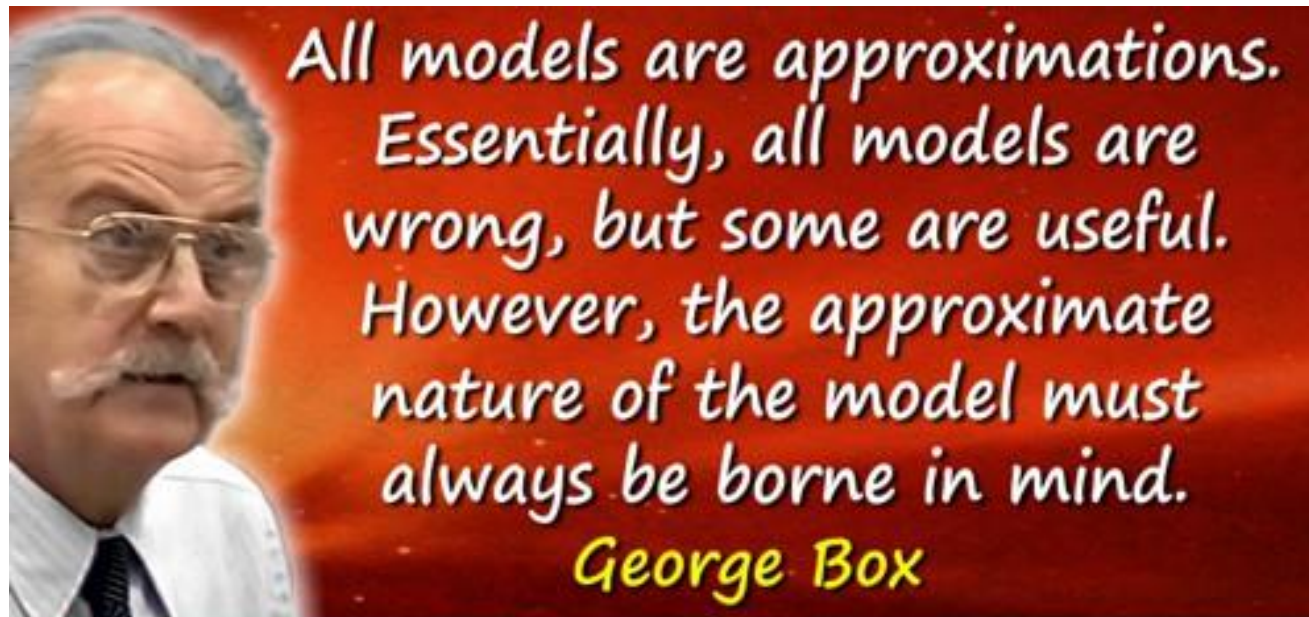
Find an **estimator** to estimate parameters of interest,

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

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

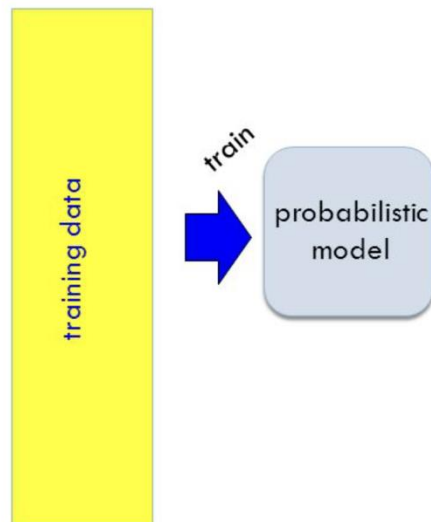
- We pose a model in the form of a probability distribution p_{θ} , with unknown **parameters of interest θ**
 - Models are found by trial and errors in different applications



More science quotes at Today in Science History todayinsci.com

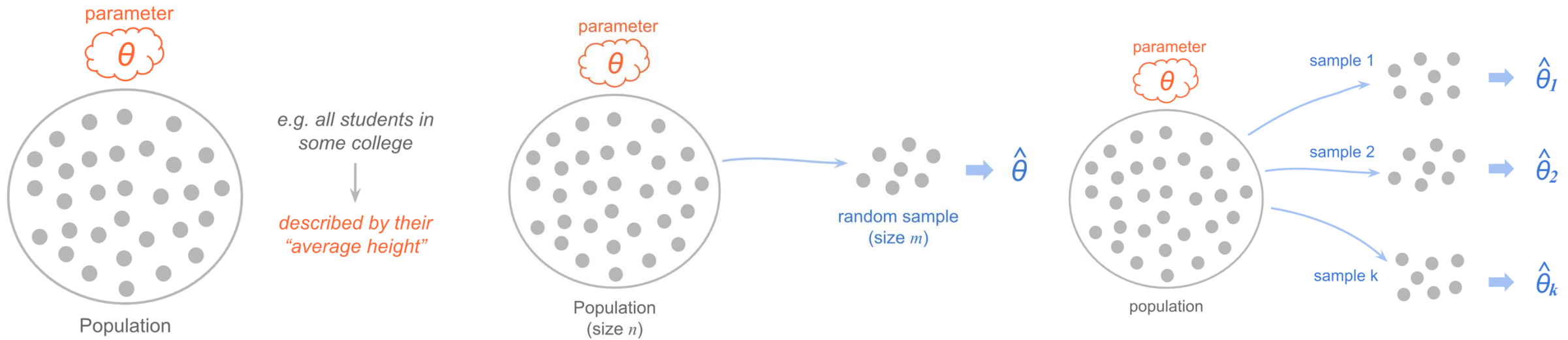
Statistical inference is sometimes called “probabilistic machine learning”:

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 $P(z; \hat{\theta})$

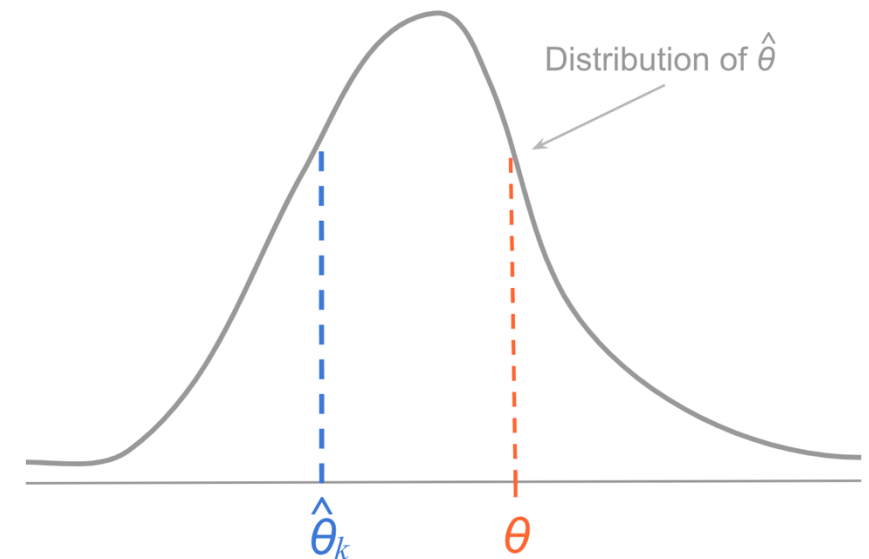


In Statistics, we mostly stop at **step 2**

Machine Learning cares more about step 3: prediction & decision



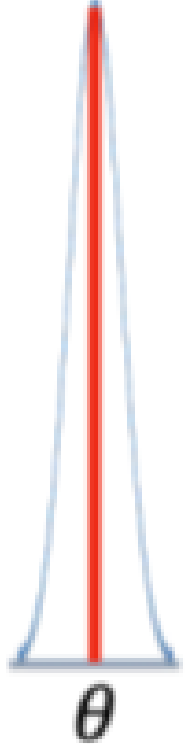
$\hat{\theta}_n$ is a random variable, it has a distribution



- 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

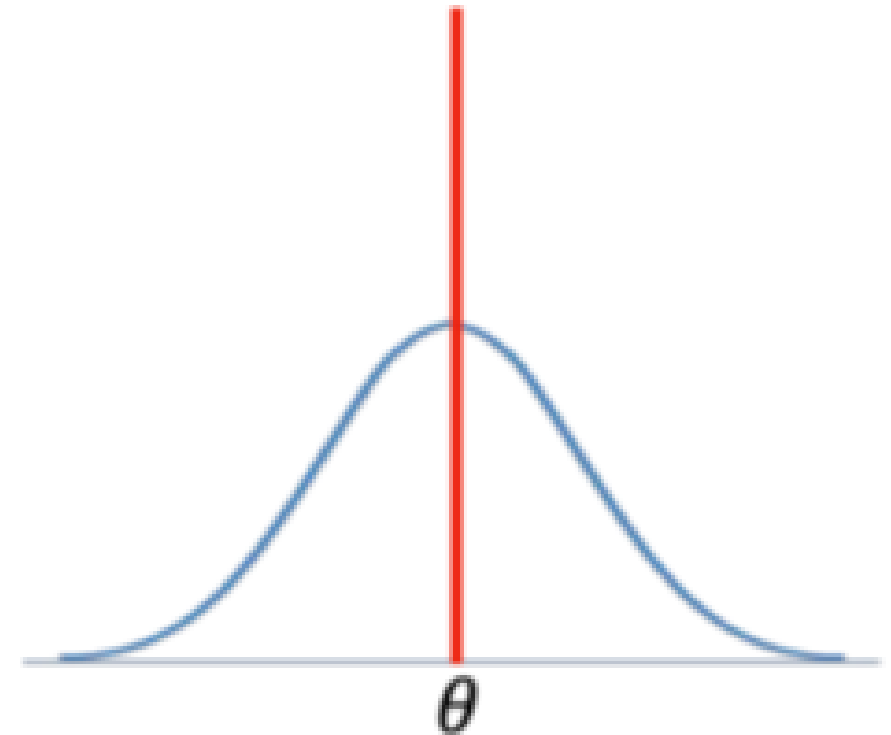
Distribution of $\hat{\theta}_n$



Good



Bad



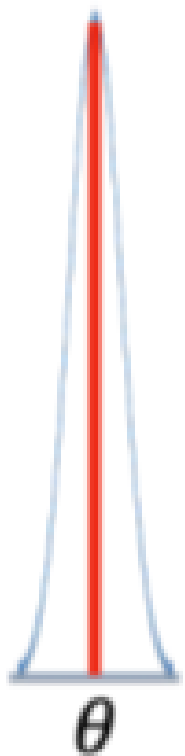
Bad

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

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

Distribution of $\hat{\theta}_n$

Low
MSE



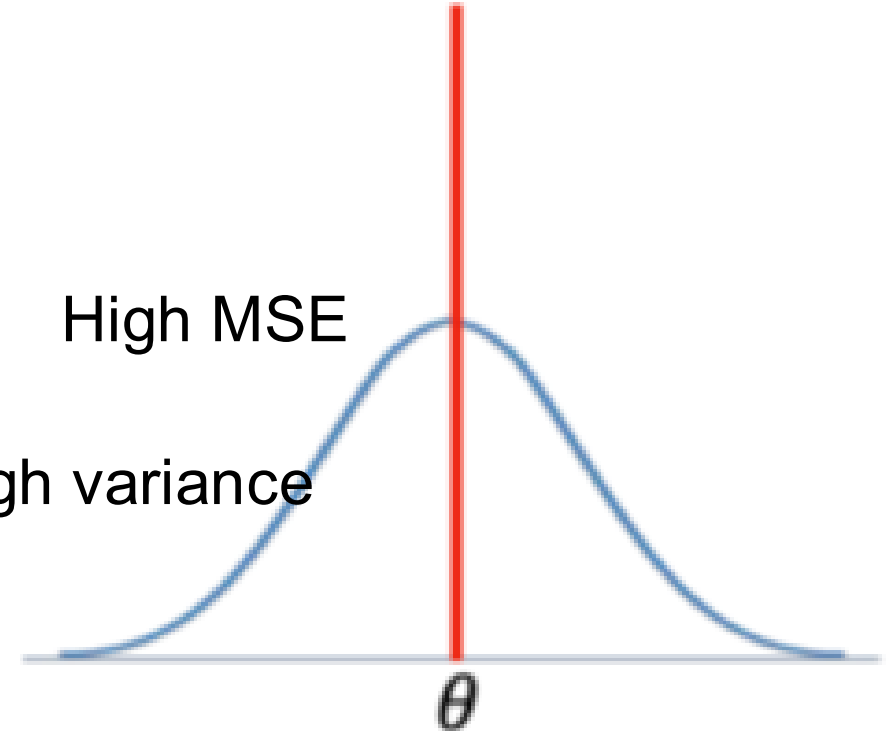
High MSE

High bias

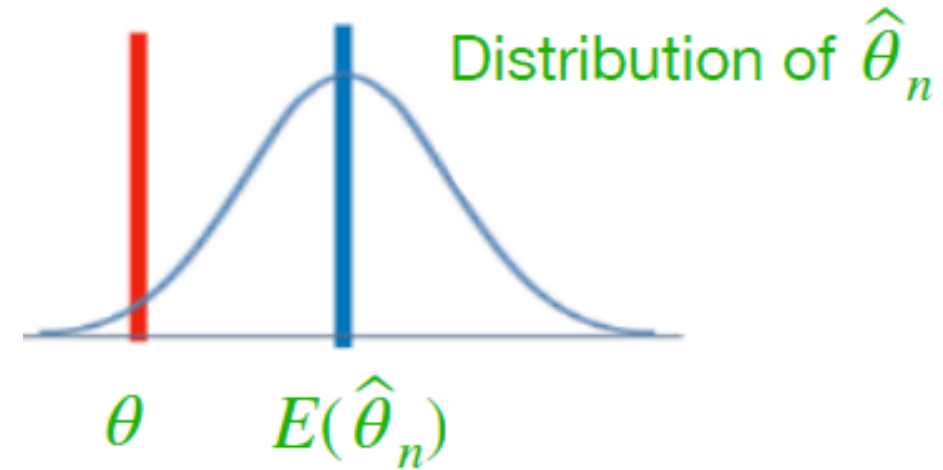


High MSE

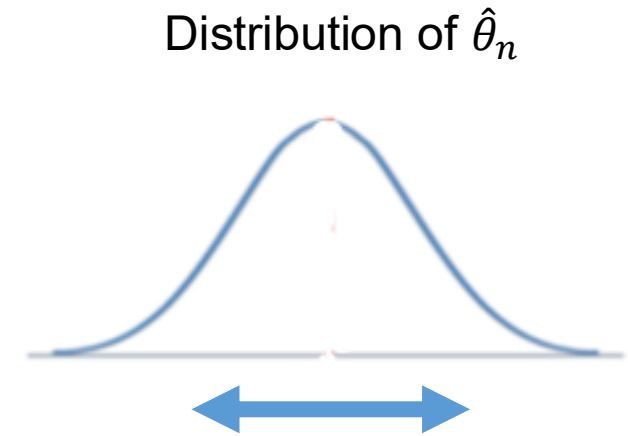
High variance



- Bias: expected overestimate of θ
- $\text{Bias}(\hat{\theta}_n) = E[\hat{\theta}_n] - \theta$
also denoted as $\mu_{\hat{\theta}_n}$
- An estimator is *unbiased* if $\text{Bias}(\hat{\theta}_n) = 0$



- Variance: how much $\hat{\theta}_n$ deviate from its mean
- $\text{Var}(\hat{\theta}_n) = E[(\hat{\theta}_n - E[\hat{\theta}_n])^2]$



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

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

Justification

$$\begin{aligned} \text{MSE} &= \text{E}\left[(\hat{\theta}_n - \mu_{\hat{\theta}_n} + \mu_{\hat{\theta}_n} - \theta)^2\right] \\ &= \text{E}\left[(\hat{\theta}_n - \mu_{\hat{\theta}_n})^2 + (\mu_{\hat{\theta}_n} - \theta)^2 + 2(\hat{\theta}_n - \mu_{\hat{\theta}_n})(\mu_{\hat{\theta}_n} - \theta)\right] \end{aligned}$$

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

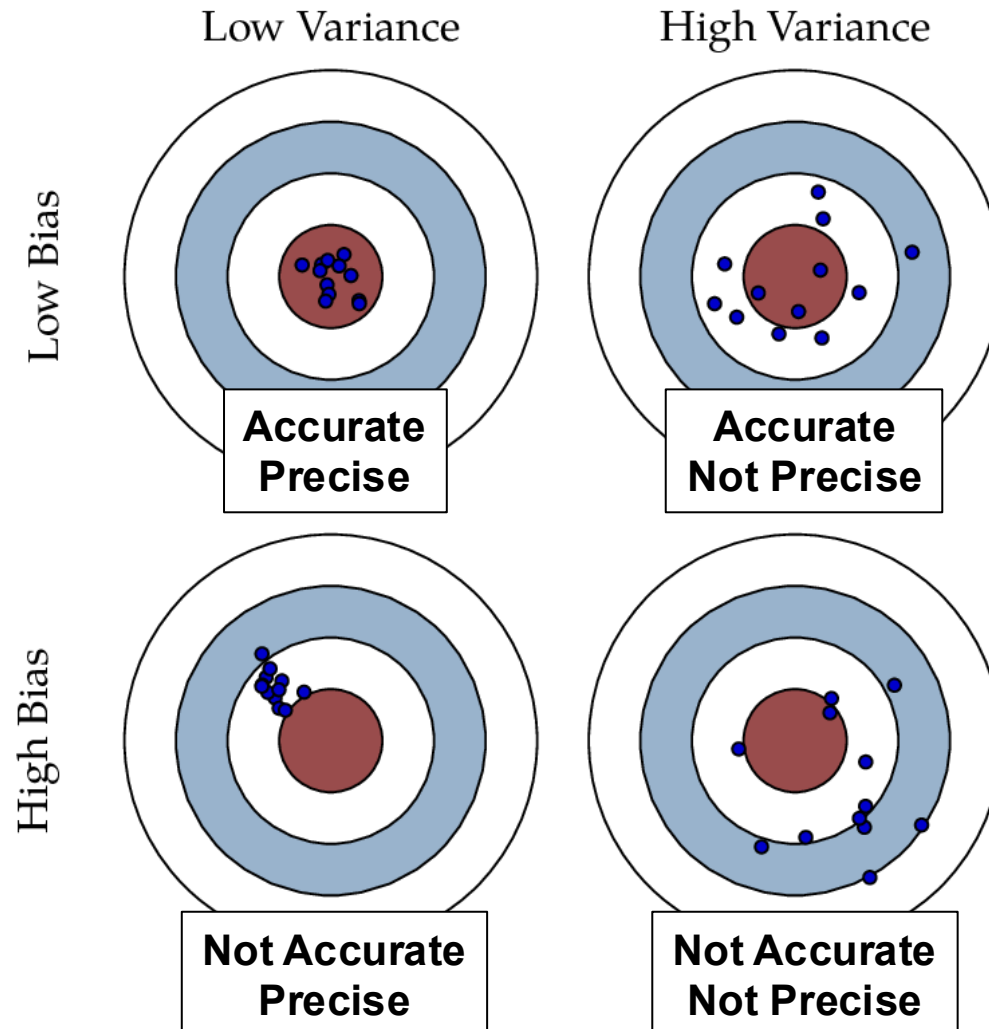
Variance

Bias

0 (why?)

a constant

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 \approx systematic error
- Variance \approx random error

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

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.

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

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

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

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



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

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

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 estimator²¹

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

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

A biased estimator

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

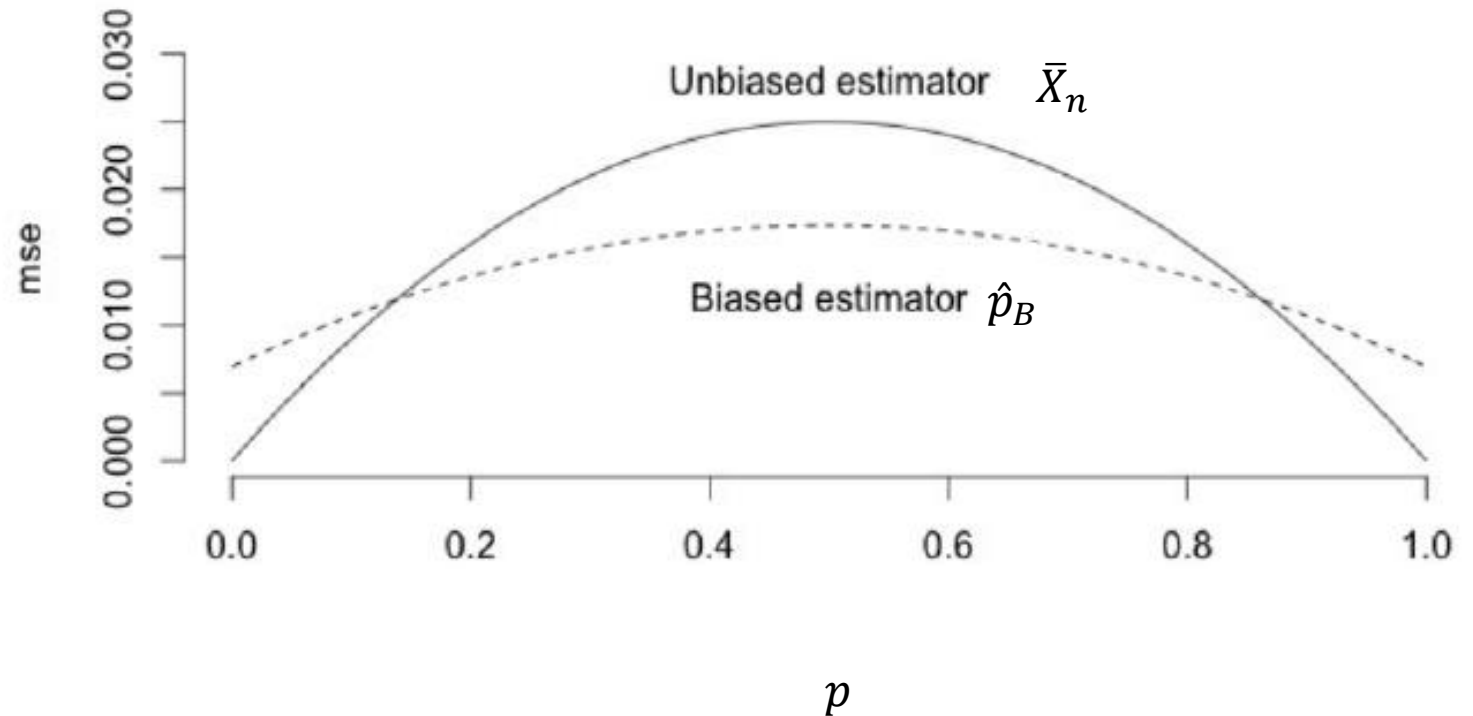
Smaller than that of
sample mean: $\frac{p(1-p)}{n}$

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

- 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...

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

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

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

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

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

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

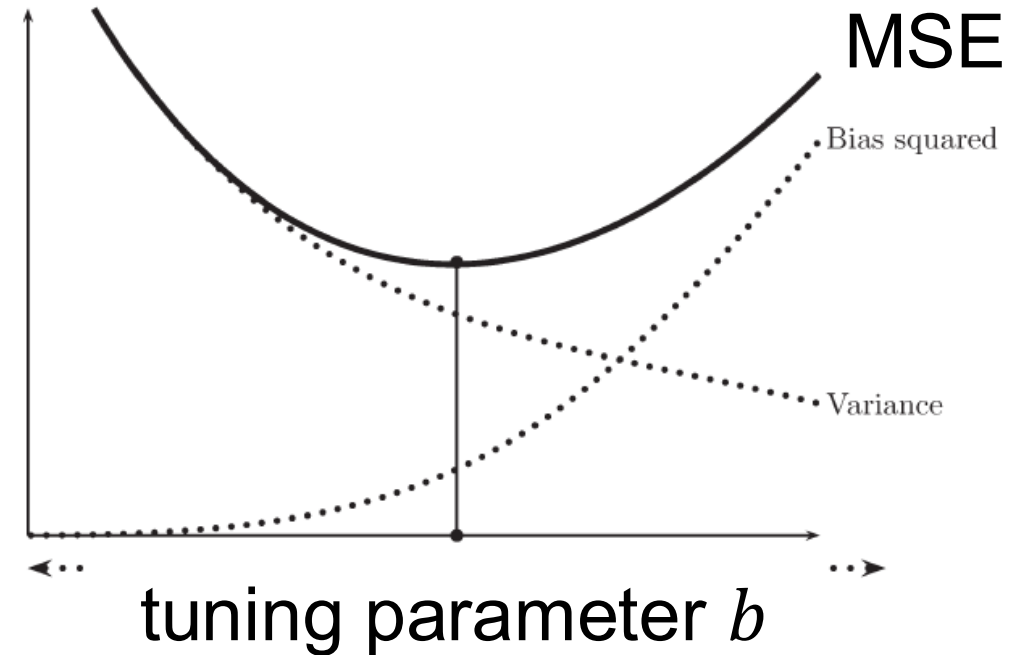
Consider a *family* of estimators $\hat{p}_b = \frac{b + \sum_i X_i}{2b + n}$ for coinflips

$b \uparrow \Rightarrow \text{bias} \uparrow, \text{variance} \downarrow$

There is some ‘sweet spot’ in choosing b

This is known as bias-variance tradeoff

akin to bias-complexity tradeoff we saw in ML

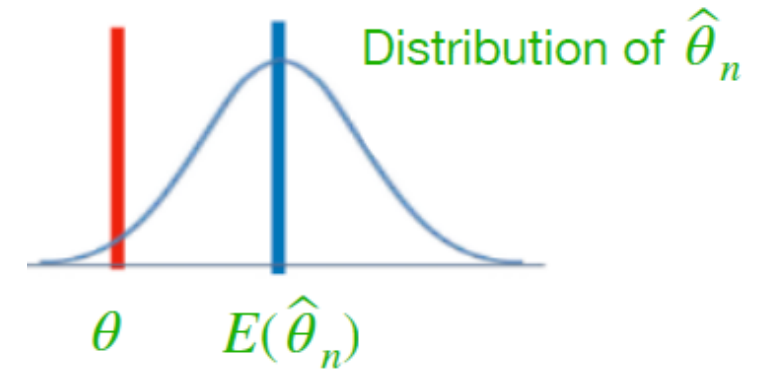


- $\text{Bias}(\hat{\theta}_n) = E[\hat{\theta}_n] - \theta$

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

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

- *An unbiased estimator is not always “better” than a biased one...*



Plug-in estimators

Property of distribution: θ

Property of samples: $\hat{\theta}_N$

Mean: $\mu = E[X] = \sum_x x f(x)$

Sample Mean: $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$

Variance: $\sigma^2 = \text{Var}[X] = E[(X - \mu)^2]$

Sample Variance: $\widehat{\sigma^2} = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$?

$$\widehat{\sigma^2} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

Correlation: $\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]}}$

Sample Correlation: $\frac{\frac{1}{N} \sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2}}$

What are the following estimators estimating?

- sample median

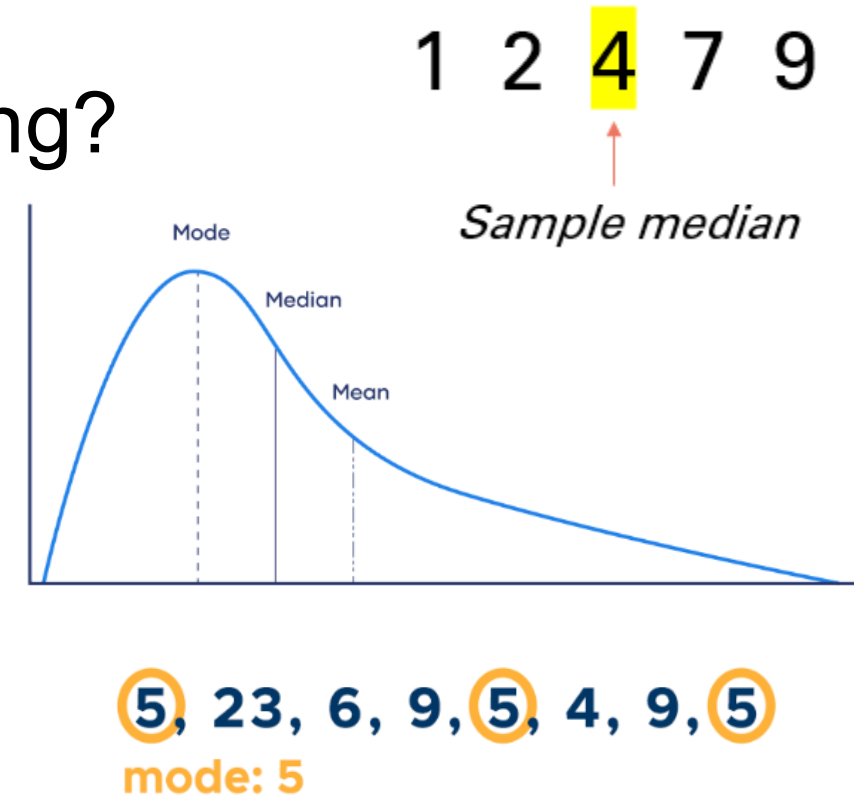
- Median of data distribution $F^{-1}\left(\frac{1}{2}\right)$

- sample mode

- Mode of data distribution $\operatorname{argmax}_x f(x)$

- sample minimum

- The minimum possible value that can be taken $\min\{x: f(x) > 0\}$



Sample variance estimator: unbiased?

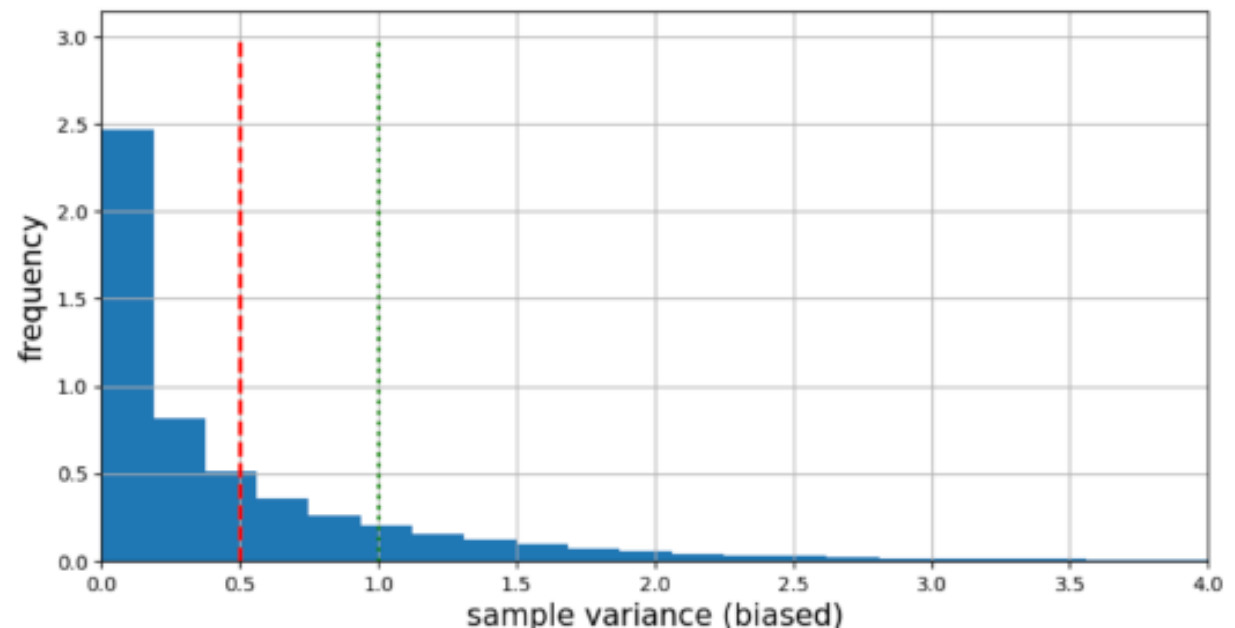
- Using $\widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$ to estimate population variance σ^2 , $N = 2$

```
n=2
s = 100000
X = np.random.normal(0,1,[n,s])
# ddof is 0(1) for dividing by n (n-1)
svar_b = np.var(X,axis=0,ddof=0)
mean_svar_b = np.mean(svar_b)
```

- Draw 2 points $\sim N(0, 1)$, repeat 100,000 times
- Computes the sample variance of the two points using the formula above
- So we have 100,000 sample variances
- Draw the distribution of the variances
- Find the mean of the distribution to be 0.5

True σ^2 : 1.0 $\widehat{\sigma}^2$'s mean: 0.5

$\widehat{\sigma}^2$ is a biased estimator of σ^2 !



- **Fact** $E[\widehat{\sigma^2}] = \frac{N-1}{N} \sigma^2$

- Bias for $\widehat{\sigma^2}$

- $E[\widehat{\sigma^2}] - \sigma^2 = \frac{N-1}{N} \sigma^2 - \sigma^2 = -\frac{1}{N} \sigma^2$

- the bias can be significant if the sample size N is small

- How can we make it unbiased?

- Scale it: $\widehat{\sigma_1^2} = \frac{N}{N-1} \widehat{\sigma^2} = \frac{N}{N-1} \cdot \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$

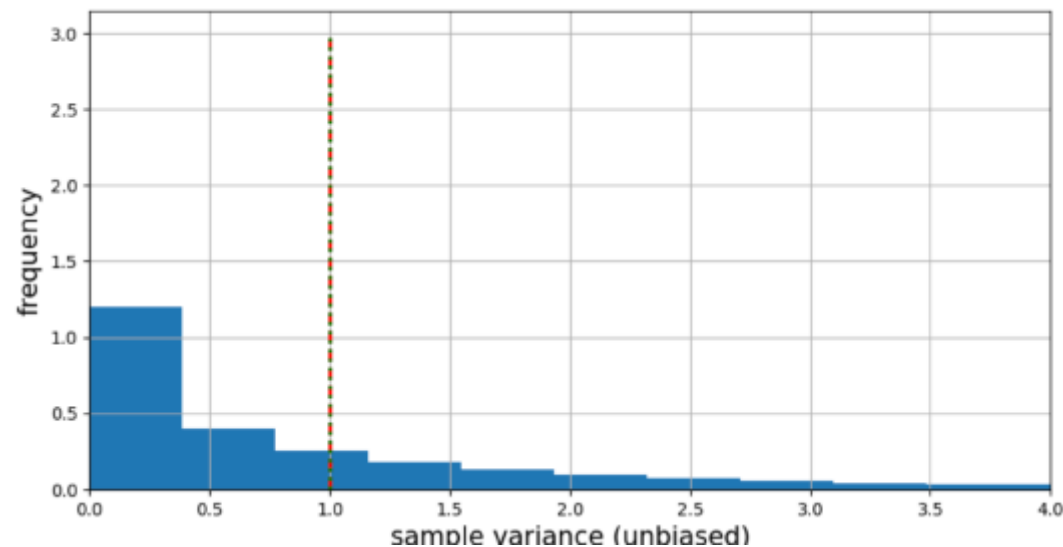
- Bias for $\widehat{\sigma_1^2} = E[\widehat{\sigma_1^2}] - \sigma^2 = \frac{N}{N-1} \widehat{\sigma^2} - \sigma^2 = \frac{N}{N-1} \cdot \frac{N-1}{N} \sigma^2 - \sigma^2 = 0$

- Using $\widehat{\sigma}_1^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$ to estimate population variance σ^2 , $N = 2$

```
n=2
s = 100000
X = np.random.normal(0,1,[n,s])
# ddof is 0(1) for dividing by n (n-1)
svar_b = np.var(X,axis=0,ddof=1)
mean_svar_b = np.mean(svar_b)
```

True σ^2 : 1.0 $\widehat{\sigma}_1^2$'s mean: 1.0

$\widehat{\sigma}_1^2$ is an unbiased estimator of σ^2

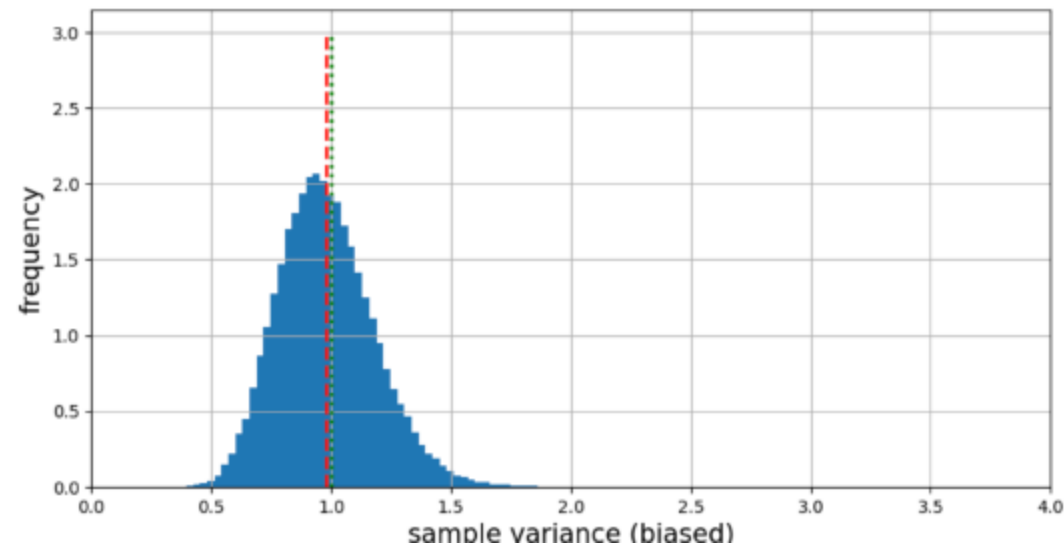


- For large N , $\widehat{\sigma}^2$ has negligible bias, and is close to $\widehat{\sigma}_1^2$

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

$$\widehat{\sigma}_1^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

```
n=50  
s = 100000  
X = np.random.normal(0,1,[n,s])  
# ddof is 0(1) for dividing by n (n-1)  
svar_b = np.var(X,axis=0,ddof=0)  
mean_svar_b = np.mean(svar_b)
```



Maximum likelihood estimators

- Likelihood: joint probability of observing this sample given model parameter

Example 4 flips of a coin -> [H, T, H, H].

What is the likelihood of observing this sample if $p = 0.75$?

$$0.75^3 0.25^1 \quad \text{Larger}$$

What is the likelihood of observing this sample if $p = 0.25$?

$$0.25^3 0.75^1 \quad \text{Smaller}$$

Larger likelihood seems to correspond to more plausible model

- Likelihood function: joint PMF / PDF of the data as a function of unknown parameter θ

- Let X_1, \dots, X_n be an iid sample with PMF / PDF $f(x; \theta)$

$$L(\theta) = \prod_{i=1}^n f(X_i; \theta) \longrightarrow \text{How well } \theta \text{ “explains” data point } X_i$$

- The maximum likelihood principle: find parameter θ that maximizes the likelihood
- Equivalently, we try to find θ that maximizes log-likelihood $\ln L(\theta)$

$$\ln L(\theta) = \sum_{i=1}^n \ln f(X_i; \theta)$$

Example $X_1, \dots, X_n \sim \text{Bernoulli}(p)$. Find the MLE for p .

x=0	x=1
1-p	p

E.g. 1, 0, 1, 1,

$$L(p) = p \cdot (1 - p) \cdot p \cdot p = p^3 \cdot (1 - p)$$

Solution

$$L(p) = \prod_{i=1}^n f(X_i; p) = p^{n_1} (1 - p)^{n_0}$$

- n_0 and n_1 are the number of 0's and 1's in the sample
- We would like to solve $\text{maximize}_{p \in [0,1]} p^{n_1} (1 - p)^{n_0}$
- Equivalently, $\text{maximize}_{p \in [0,1]} n_1 \ln p + n_0 \ln(1 - p)$

Math Interlude: optimization problems

A constrained optimization problem:

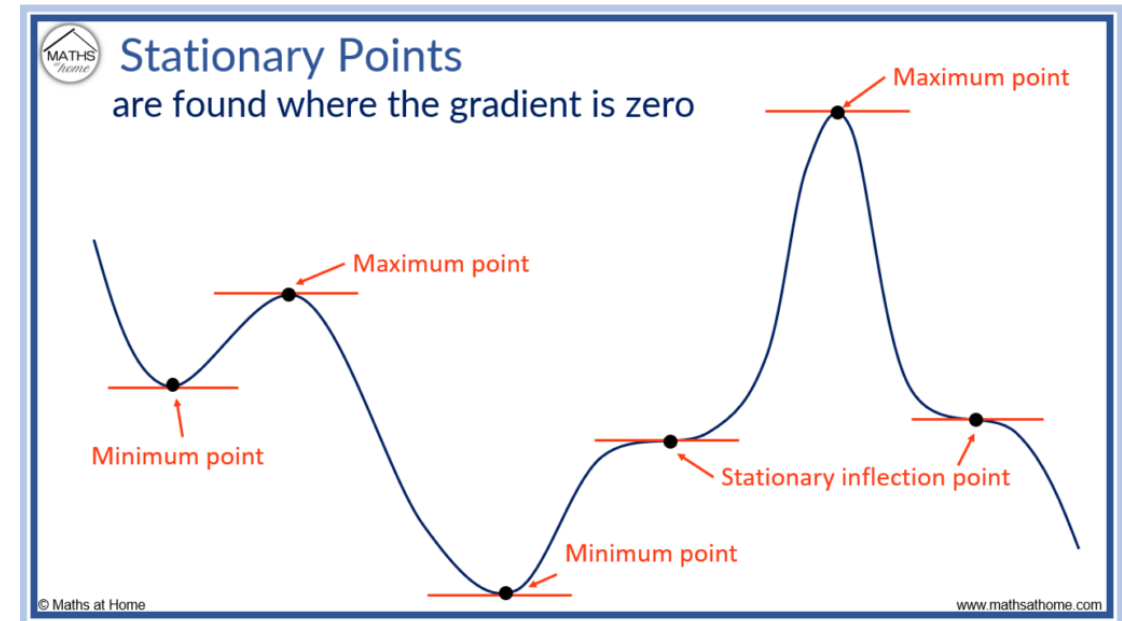
$$\text{maximize}_{p \in [0,1]} n_1 \ln p + n_0 \ln(1 - p)$$

p : **Optimization variables**

$p \in [0,1]$: **constraint**

Note Setting the objective's derivative to zero and solve for p gives a *stationary point*, but

- It may be local maximum, or even *local minimum*
- It may fall out of constraint set
- We recommend always plotting the objective function to check

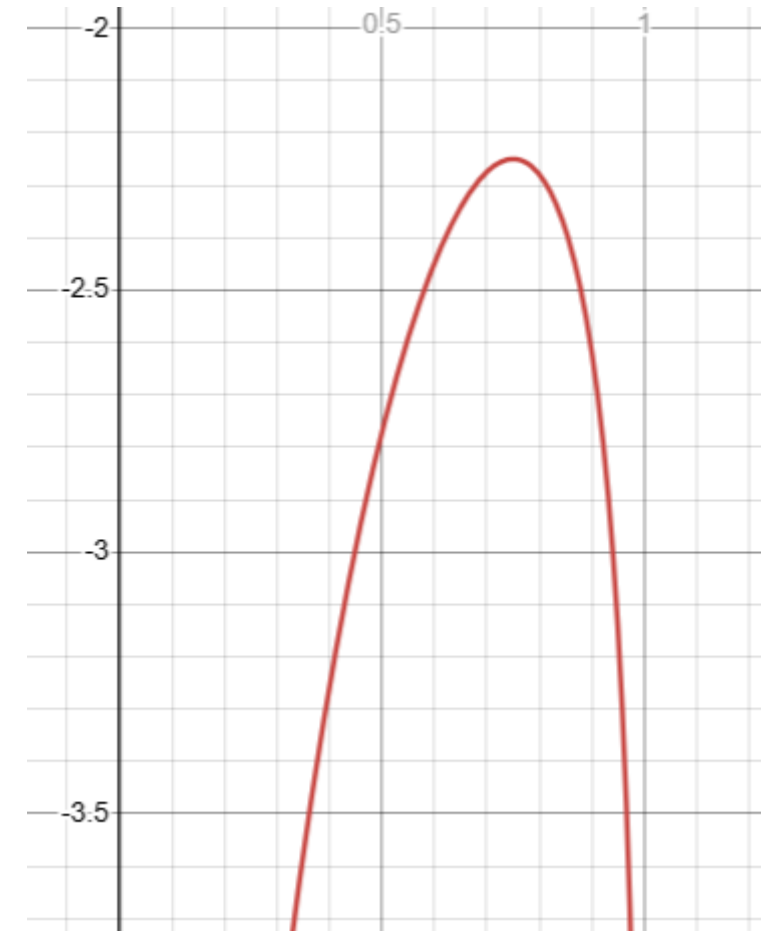


$$\text{maximize}_{p \in [0,1]} n_1 \ln p + n_0 \ln(1 - p)$$

- E.g. when $n_0 = 1, n_1 = 3$, we have:
- Its global maximizer indeed lie in its only stationary point

Stationary point can be found by

$$\frac{n_1}{p} - \frac{n_0}{1-p} = 0 \quad \Rightarrow \quad p = \frac{n_1}{n_1+n_0} = \frac{n_1}{n} = \text{sample mean}$$



Example Assume that UA students' heights (in centimeters) follow $N(\mu, 8^2)$, and we observe 4 students' heights:

163, 171, 179, 167

Find the maximum likelihood estimator for μ

Solution

Step 1: write down the log-likelihood function

$$\ln L(\mu) = \sum_{i=1}^n \ln f(x_i; \mu) \quad f(x_i; \mu) = \frac{1}{\sqrt{2\pi}8^2} \exp\left(-\frac{(x_i - \mu)^2}{2 \times 8^2}\right)$$

the sum has 4 terms -- e.g. the first term is $\ln \frac{1}{\sqrt{2\pi}8^2} - \frac{(163 - \mu)^2}{2 \times 8^2}$

- **Step 2: simplify the log-likelihood function**

4 samples: 163, 171, 179, 167

$$\begin{aligned}\ln L(\mu) &= \sum_{i=1}^n \left(\ln \frac{1}{\sqrt{2\pi}8^2} - \frac{(x_i - \mu)^2}{2 \times 8^2} \right) \\ &= -\frac{1}{128} \sum_{i=1}^4 (x_i - \mu)^2 - 11.99\end{aligned}$$

- **Step 3: find μ that maximizes log-likelihood:**

Fact: the μ that minimizes $\sum_{i=1}^4 (x_i - \mu)^2$ is $\mu = \bar{x}$

the μ that maximizes $L(\mu)$ is $\bar{x} = \frac{163+171+179+167}{4} = 170$

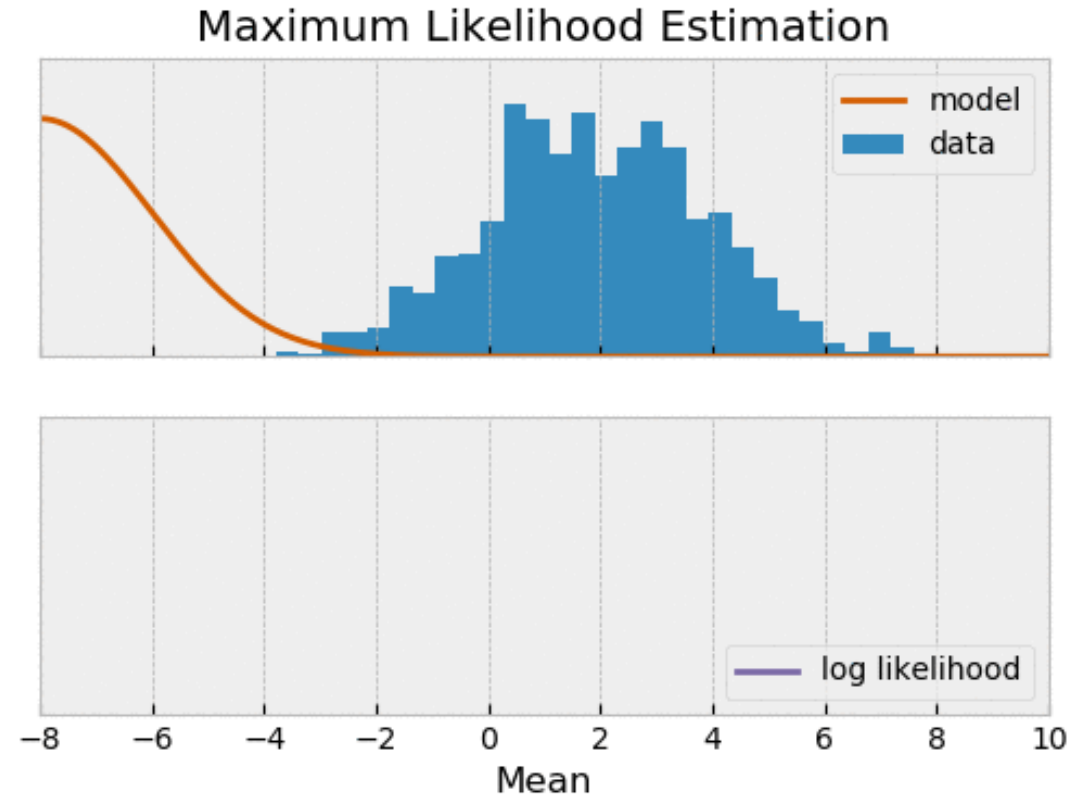
Summary given the data we have, we estimate that UA students' heights follow $N(170, 8^2)$

How would you use it to predict an unseen UA student's height?
perhaps 170cm is a decent guess..

General Fact Fixed σ (e.g. $\sigma = 8$). Assume samples x_1, \dots, x_n are drawn from $N(\mu, \sigma^2)$. Then the MLE for μ is sample mean

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

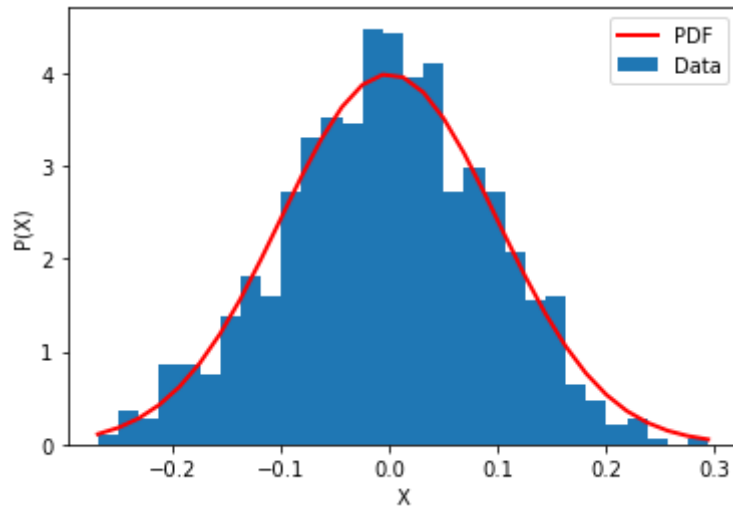
Among all $N(\mu, \sigma^2)$'s, $\mu = \bar{x}$ has the highest likelihood



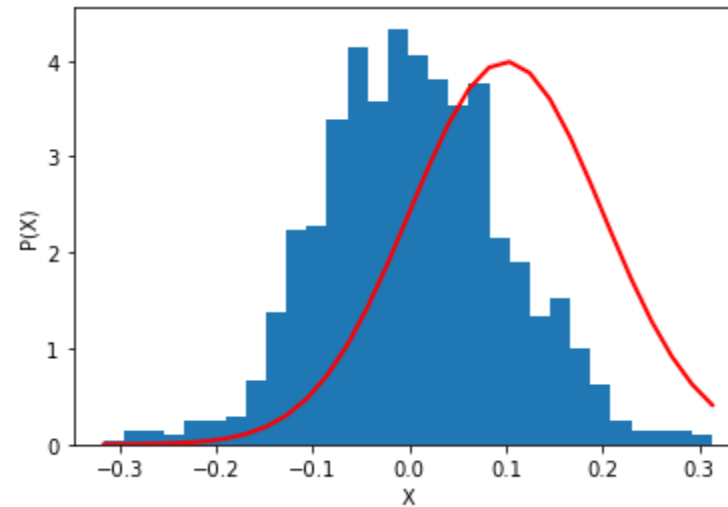
Suppose we observe n data points from a Gaussian model $N(\mu, \sigma^2)$ and wish to estimate **both μ and σ**

Say we only need to choose from the following three Gaussians...

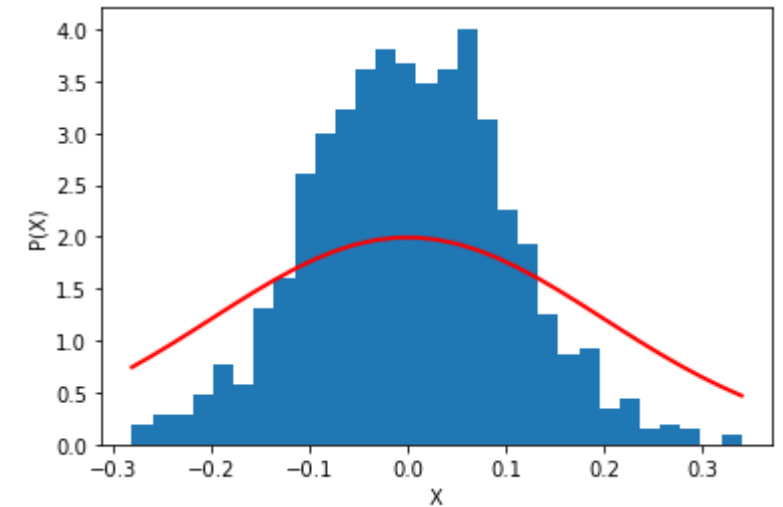
High Likelihood



Low Likelihood (mean)



Low Likelihood (variance)



Here, $L(\mu, \sigma^2) = \prod_{i=1}^n f(x_i; \mu, \sigma)$ $f(x_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$

Fact Assume samples x_1, \dots, x_n are drawn from $N(\mu, \sigma^2)$. Then the MLE for μ, σ is given by:

$$\mu = \bar{x} \text{ (sample mean)}$$

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \text{ (sample standard deviation)}$$

Example Assume that UA students' weights (in kg) follow a Gaussian, and we observe 4 students' weights: 60, 65, 70, 75

$$\text{The MLE } \mu = \frac{60+65+70+75}{4} = 67.5,$$

$$\text{MLE } \sigma = \dots = 5.6$$

Therefore our estimate of UA students' weights $\sim N(67.5, 5.6^2)$

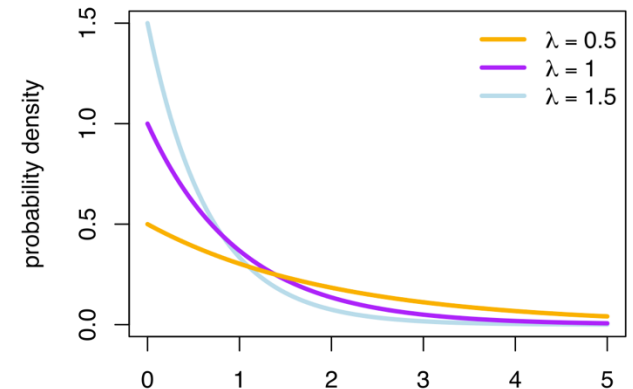
Wait time at the barbershop: Suppose you go to a barbershop every quarter. You want to be able to predict the waiting time. You have collected 4 data points (in minutes) from last year:

3, 2, 6, 5

Suppose we model the waiting time using an exponential distribution
Exponential(λ):

$$f(x) = \lambda e^{-\lambda x}$$

- Find the maximum likelihood estimator for λ
- How would you use this to predict your next waiting time?



Step 1: write down the log-likelihood function

$$\ln L(\lambda) = \sum_{i=1}^n \ln f(x_i; \lambda) \qquad f(x; \lambda) = \lambda e^{-\lambda x}$$

Step 2: simplify the log-likelihood function

$$\ln L(\lambda) = \sum_{i=1}^n (\ln \lambda - \lambda x_i) = n \ln \lambda - \lambda \sum_{i=1}^n x_i$$

for our dataset, this is

Dataset: 3, 2, 6, 5

$$4 \ln \lambda - \lambda(3 + 2 + 6 + 5) = 4 \ln \lambda - 16 \lambda$$

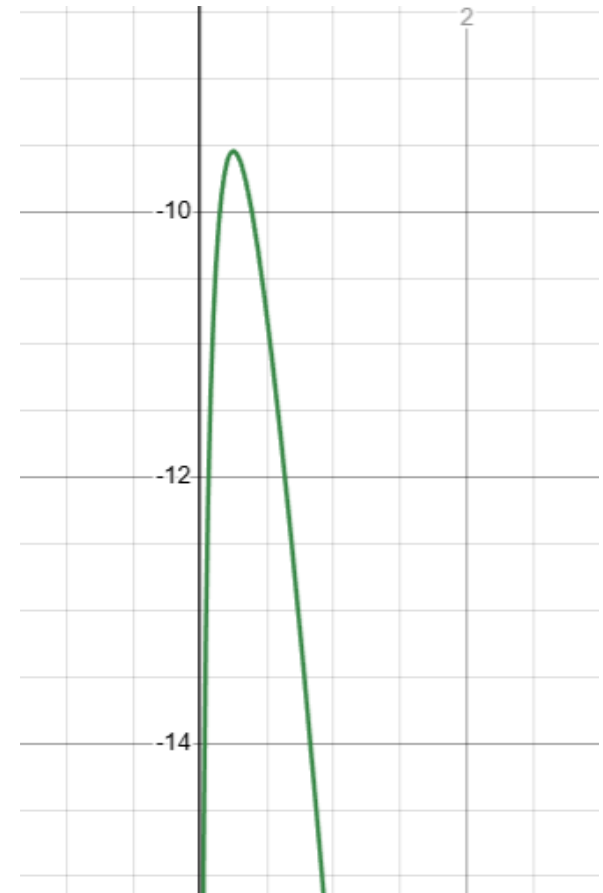
- **Step 3: maximize the log-likelihood function**

Maximize $L(\lambda) = 4 \ln \lambda - 16 \lambda$

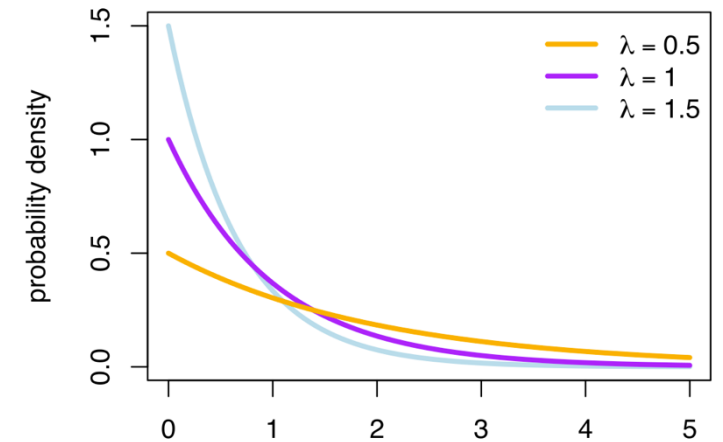
It has only one stationary point,
which corresponds to its maximum

can be found by solving $L'(\lambda) = 0$

$$\frac{4}{\lambda} - 16 = 0 \quad \Rightarrow \quad \lambda = \frac{1}{4}$$



Summary given the data, we estimate the waiting time to follow the Exponential ($\lambda = \frac{1}{4}$) distribution



How would you use this to predict your next waiting time?

We can use the mean of our learned distribution:

$$\frac{1}{\lambda} = 4 \text{ (minutes)} \quad f(x) = \frac{1}{4} e^{-\frac{x}{4}}$$