



Computer  
Science

# CSC380: Principles of Data Science

## **Statistics 3**

Xinchen Yu

# Review: Maximum Likelihood Estimation

Suppose  $x_i \sim p(x; \theta)$ , the **joint probability** over N i.i.d  $x_1, \dots, x_N$

$$p(x_1, \dots, x_N; \theta) = \prod_{i=1}^N p(x_i; \theta)$$

**Maximum Likelihood Estimator (MLE)** as the name suggests, maximizes the likelihood function.

$$\hat{\theta}^{\text{MLE}} = \arg \max_{\theta} \mathcal{L}_N(\theta) = \prod_{i=1}^N p(x_i; \theta)$$

**Log Likelihood Maximum**

$$\hat{\theta}^{\text{MLE}} = \arg \max_{\theta} \log \mathcal{L}_N(\theta) = \sum_{i=1}^N \log p(x_i; \theta)$$

Finding the MLE:

1. closed-form
2. iterative methods

Q: is Maximum likelihood estimator consistent? efficient?

# What we will learn

- Using MLE we can get
  - sample mean to estimate true mean (e.g., bernoulli, gaussian)
  - sample variance to estimate true variance (e.g., gaussian)
- How can we trust maximum likelihood estimators? Do they have the desirable properties of an estimator?
  - sample mean: biased or unbiased?
  - sample variance: biased or unbiased?
    - biased version
    - unbiased version
    - which one is better?

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

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

- ❖ **Efficiency** Mean squared error should be small

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

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

- ❖ **Unbiasedness** It should be unbiased (not always true !)

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

1) The MLE is a **consistent** estimator:

$$\lim_{n \rightarrow \infty} \hat{\theta}_n^{\text{MLE}} \xrightarrow{P} \theta_*$$

Roughly, converges to the true value.

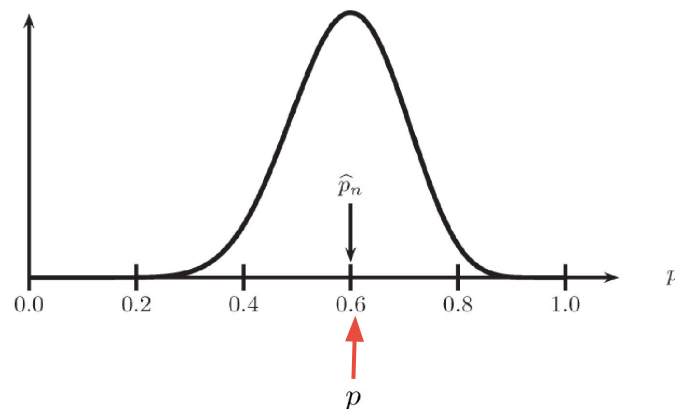
2) The MLE is **efficient**: roughly, has the lowest mean squared error among all consistent estimators.

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

3) The MLE is **Normal**: roughly, the estimator (which is a random variable) approaches a Normal distribution.

Note: we did not say MLE is unbiased

3) The MLE is **Normal**: roughly, the estimator (which is a random variable) approaches a Normal distribution.



- We pick  $k$  different samples (each sample has  $N$  i.i.d observations)
- We pose a model with unknown parameter
- Get MLE estimation for the parameter (a total of  $k$  estimators)
- The distribution of  $k$  estimators is roughly normal distribution
  - Expectation
  - Variance

## Sample Mean: Expectation and Variance

# Review: Bernoulli Expectation and Variance

**Bernoulli** A.k.a. the **coinflip** distribution on binary RVs  $X \in \{0, 1\}$

$$p(X) = \pi^X (1 - \pi)^{(1-X)}$$

$$\text{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

Where  $\pi$  is the probability of **success** (i.e., heads), and also the mean

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

$$\text{Var}[X] = \pi(1 - \pi)$$

$$\mathbf{E}[X^2] = \pi \cdot 1^2 + (1 - \pi) \cdot 0^2 = \pi$$

$$\text{Var}[X] = \pi - \pi^2$$

We knew MLE for Bernoulli is sample mean.  
Q: for sample mean, what's  $\mathbf{E}[X]$  and  $\text{Var}[X]$ ?





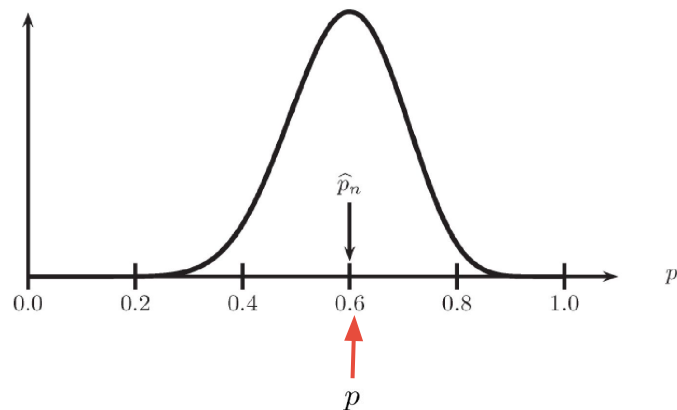
*Recall: An estimator  $\hat{\theta}$  is a RV (Random Variable).*

**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** What is the expected value of  $\hat{p}$ ?

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)

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

quiz candidate

$$\begin{aligned}\text{Var}(\hat{p}) &= \text{Var}\left(\frac{1}{N} \sum_i X_i\right) \stackrel{(a)}{=} \frac{1}{N^2} \text{Var}\left(\sum_i X_i\right) \stackrel{(b)}{=} \frac{1}{N^2} \sum_i \text{Var}(X_i) \\ &\stackrel{(c)}{=} \frac{1}{N^2} \sum_i p(1-p) = \frac{1}{N} p(1-p) = \frac{1}{N} \text{Var}(X)\end{aligned}$$

(a)  $\text{Var}(cX) = c^2 \text{Var}(X)$

(b) Independent RVs

(c)  $\text{Var}(X) = p(1-p)$  for Bernoulli

**In General** Variance of sample mean  $\bar{X}$  for RV with variance  $\sigma^2$

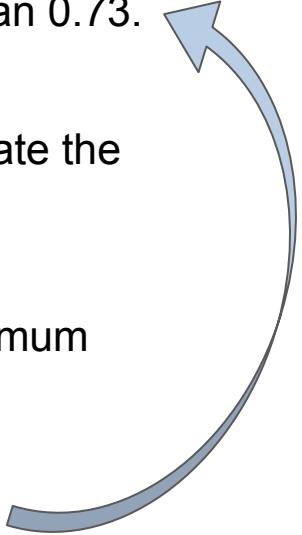
**STDEV of sample mean  
decreases as  $1/\sqrt{N}$**

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{N}$$

**Decreases linearly with  
number of samples N**

# All Facts about Sample Mean

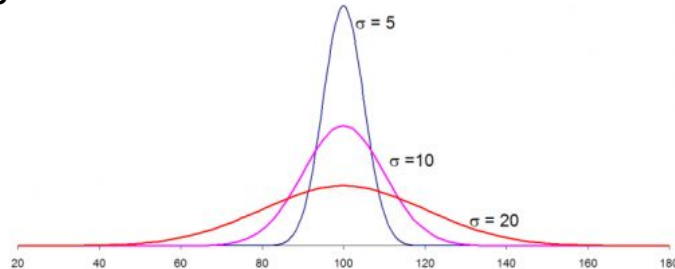
Experiment Flip a coin 100 times and observe 73 heads, 27 tails

- We don't know the coin bias. By intuition, we guess coin bias is sample mean 0.73.
  - We are told that maximum likelihood estimation is a method that can estimate the parameter of an assumed probability distribution.
  - So we pose a model of bernoulli, and calculate the estimator that can maximum the log likelihood function.
  - We find the maximum likelihood estimator is sample mean = our intuition!
- 

# All Facts about Sample Mean

Experiment Flip a coin 100 times and observe 73 heads, 27 tails

- If we repeatedly flip a coin 100 times ( $N=100$ ), say 1000 trails (1000 samples). We will get 1000 sample means. So sample mean is also a RV. It has a distribution.
- Pile 1000 sample means up, we get a distribution (roughly normal). The mean of the distribution (expectation) = true coin bias.
- If we flip a coin 10,000 times ( $N=10,000$ ), repeat for 1000 trails (1000 samples). The variance of the distribution is very small. We can trust the sample mean more when estimating true coin bias



## Sample Variance: Expectation

# Unbiasedness of the Sample Variance?

Recall: Sample mean is an unbiased estimator for the true mean.

***How about the sample variance?***

**Ex.** Let  $X_1, \dots, X_N$  be drawn (iid) from any distribution with  $\text{Var}(X) = \sigma^2$  and,

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

Then the sample variance is a **biased estimator**.

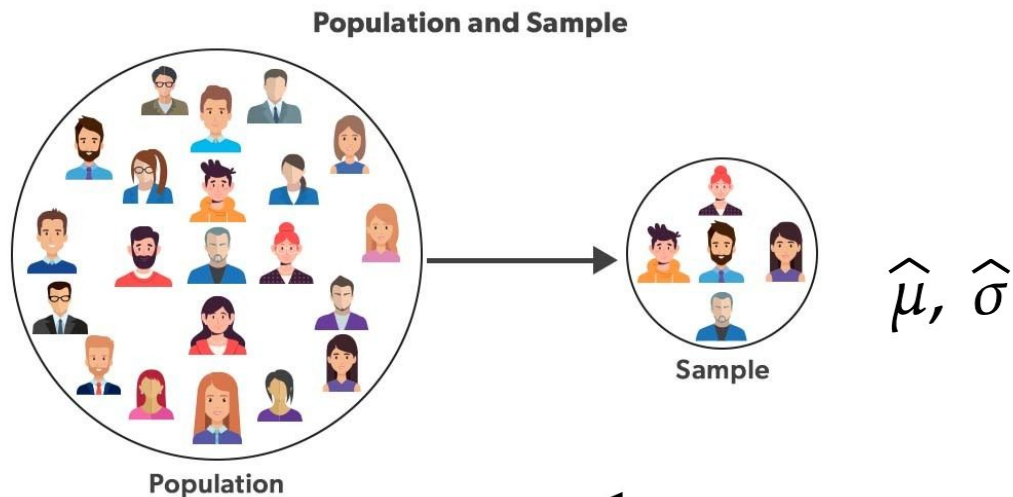
Source of bias:  
plug-in mean  
estimate

$$\mathbf{E}[\hat{\sigma}^2] = \frac{1}{N} \sum_i \mathbf{E}[(X_i - \hat{\mu})^2] = \text{boring algebra} = \frac{N-1}{N} \sigma^2 \quad \text{tends to underestimate}$$

Q: is this estimator consistent or not? Consistent!

Correcting bias yields unbiased variance estimator:

$$\hat{\sigma}_{\text{unbiased}}^2 = \frac{N}{N-1} \hat{\sigma}^2 = \frac{1}{N-1} \sum_i (X_i - \hat{\mu})^2 \quad E[\hat{\sigma}_{\text{unbiased}}^2] = \sigma^2$$



Pose a Gaussian model  
with unknown parameters:

$\mu, \sigma$

$$\hat{\mu} = \frac{1}{n} \sum_i x_i$$

Estimate parameter mean:  
sample mean --- unbiased,  
consistent, efficient.

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \hat{\mu})^2 \quad \text{MLE: Biased}$$

$$\hat{\sigma}_{unbiased}^2 = \frac{1}{n-1} \sum_i (x_i - \hat{\mu})^2 \quad \text{Unbiased}$$

Estimate parameter variance: biased or unbiased?

# Numpy Background

- Often, you have a matrix of data: e.g., movie review score

User \ Movie	Inception	Jurassic park	Batman
A	5	2	3
B	1	4	2
C	4	3	3
D	1	2	3

Numpy arrays can be 2d

```
A = np.array([[1,2,3],[4,5,6]])
```

```
A[0,1]
```

```
⇒ 2
```

```
mean(A,0)
```

```
⇒ array([2.5, 3.5, 4.5])
```

```
mean(A,1)
```

```
⇒ array([2., 5.])
```

means  $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$

access A[0,1] means 1<sup>st</sup> row, 2<sup>nd</sup> column

computes average for each column

computes average for each row

var(A,0), var(A,1) works the same way!



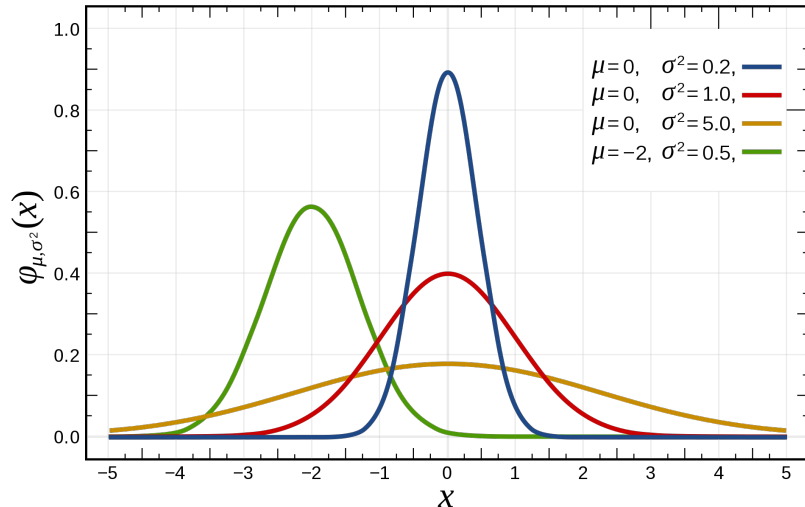
# Numpy Background

`random.randn(d0, d1, ..., dn)`

Return a sample (or samples) from the “**standard normal**” distribution (red line):

- mean is 0
- variance is 1

2 by 4 array of samples from the normal distribution with mean 3 and standard deviation 2.5:



```
>>> 3 + 2.5 * np.random.randn(2, 4)
array([[ -4.49401501,  4.00950034, -1.81814867,  7.29718677], # random
       [ 0.39924804,  4.68456316,  4.99394529,  4.84057254]]) # random
```

# MSE of biased vs unbiased version

Let's compare MSE of biased and unbiased sample variance:

- For each sample, generate 5 numbers from standard normal distribution;
- Repeat for 10,000 times, so we get 10,000 samples;
- For each sample, get
  - biased sample variance
  - unbiased sample variance
- So we have 10,000 biased/unbiased sample variances;
- For the 10,000 biased sample variances, calculate MSE

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

$$\hat{\sigma}_{unbiased}^2 = \frac{1}{n-1} \sum_i (x_i - \hat{\mu})^2$$

$$\begin{pmatrix} \hat{\sigma}_1 : -0.28, -0.74, 0.60, -0.31, 0.49 \\ \hat{\sigma}_2 : -0.31, -0.08, 0.58, 0.09, -0.53 \\ \dots\dots\dots \\ \hat{\sigma}_{10000} : 0.09, -0.82, 1.29, -1.08, -0.17 \end{pmatrix}$$

$$MSE(\hat{\sigma}) = E[(\hat{\sigma} - \sigma)^2]$$

↑  
1

**Task:** Compare the **MSE** (mean squared error) of the two variance estimators for  $N=5$ .

```
import numpy as np
import numpy.random as ra
X = ra.randn(10_000,5)    # 10k by 5 matrix of  $\mathcal{N}(0,1)$  => 10k random trials
```

```
np.mean((var(X,1,ddof=0) - 1)**2)    ddof=0 uses 1/N
⇒ 0.36310526687176103
```

```
np.mean((var(X,1,ddof=1) - 1)**2)    ddof=1 uses 1/(N-1)
⇒ 0.5071783438808787
```

biased version is more accurate! (but recall that it will underestimate)

*There is a trade off between bias and variance!!*

*Is an unbiased estimator “better” than a biased one? It depends...*

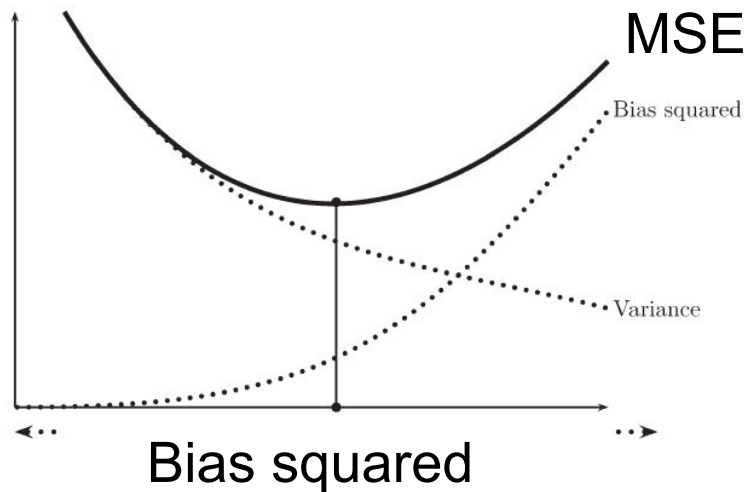
Evaluate the quality of estimate  $\hat{\theta}$  using **mean squared error**,

$$\text{MSE}(\hat{\theta}) = \mathbf{E} \left[ (\hat{\theta} - \theta)^2 \right] = \text{bias}^2(\hat{\theta}) + \mathbf{Var}(\hat{\theta})$$

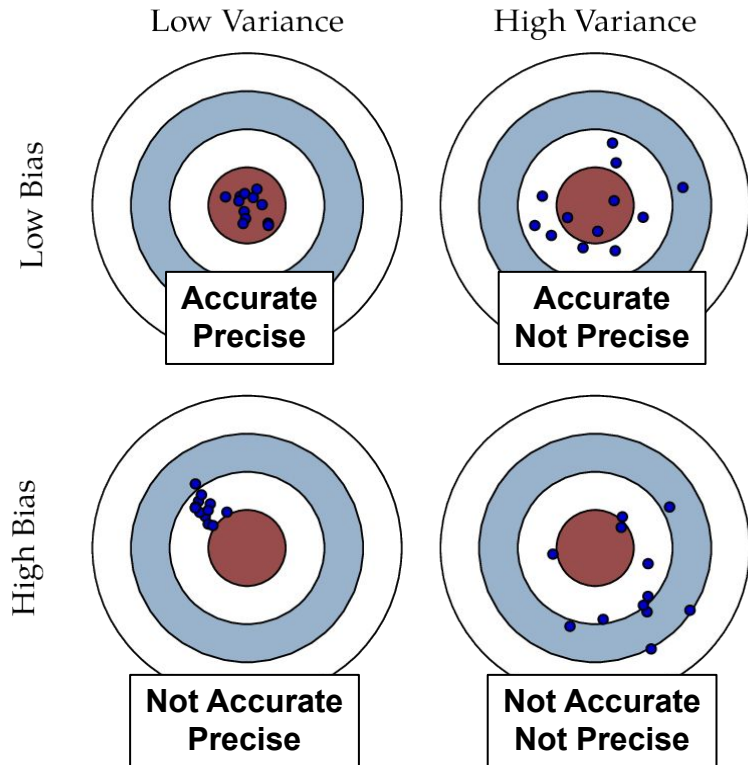
- MSE for unbiased estimators is just,

$$\text{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta})$$

- Bias-variance is fundamental tradeoff in statistical estimation
- MSE increases as **square** of bias
- Biased estimator can be more accurate than an unbiased one.



*Suppose an archer takes multiple shots at a target...*



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

- Bias: distance from the center of target
- Variance: distance from the center of multiple shots

MSE: MLE < Sample variance

- higher bias and lower var can be more efficient than lower bias and higher var.

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

quiz candidate

*Compare the results of two coin flip experiments...*

Experiment 1 Flip 100 times and observe 73 heads, 27 tails

Experiment 2 Flip 1,000 times and observe 730 heads, 270 tails

Question The MLE estimate of coin bias for both experiments is equivalent  $\hat{\theta} = 0.73$ . Which should we trust more? Why?

Answer: biases are the same (MLE use sample mean and therefore unbiased). Variance is smaller for experiment 2 (larger N). The estimator in Experiment 2 has smaller MSE.

