



## CSC380: Principles of Data Science

### Statistics 3

Xinchen Yu

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

1

## 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

2

## Maximum Likelihood Estimator Properties

3

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.

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

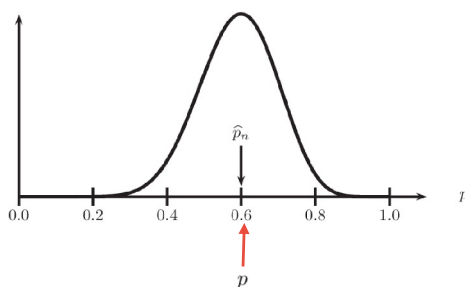
3) The MLE is **Normal**: roughly, the estimator (which is a random variable) approaches a Normal distribution.

3

## Maximum Likelihood Estimator Properties

4

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

Q: for sample mean,  
what's  $E[X]$  and  $\text{Var}[X]$ ?

4

5

## Sample Mean: Expectation and Variance

5

## Review: Bernoulli Expectation and Variance

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

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

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

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

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

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

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

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



6

## Expectation of the Sample Mean

7

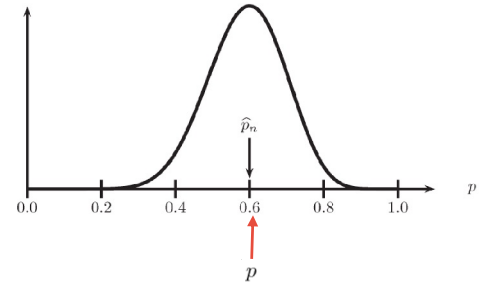
*Recall: An estimator  $\hat{\theta}$  is a RV (Random Variable).*

**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)

7

## Variance of the Sample Mean

8

**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

8

## 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!

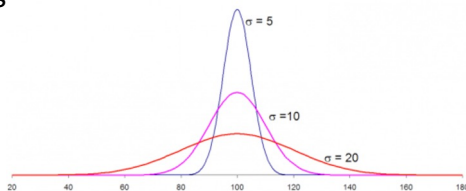


9

## 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



10

11

## Sample Variance: Expectation

11

## Unbiasedness of the Sample Variance?

12

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:

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

12

## Sample Variance estimator vs MLE estimator

13

**Population and Sample**

Population      Sample       $\hat{\mu}, \hat{\sigma}$

Pose a Gaussian model with unknown parameters:  $\mu, \sigma$

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

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

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

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

For parameter variance: biased or unbiased?

13

## Numpy Background

14

- 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!

14

## More on Unbiased Estimator

15

**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
```

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

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

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

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

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

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

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

15

## Bias-Variance Tradeoff

16

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

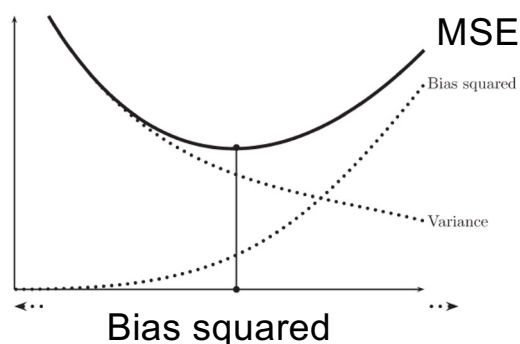
Evaluate the quality of estimate  $\hat{\theta}$  using **mean squared error**,

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

- MSE for unbiased estimators is just,

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

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



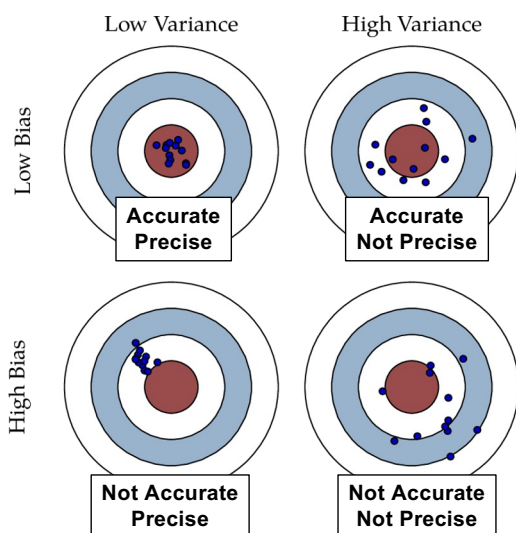
16



## Bias-Variance Tradeoff

17

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.

17

## Bias-Variance Decomposition

18

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

quiz candidate

18

## Intuition Check

19

*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.



19