CSE 312

Foundations of Computing II

Lecture 23: Maximum Likelihood Estimation (cont.)

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Anna R. Karlin

Slide Credit: Based on Stefano Tessaro's slides for 312 19au
incorporating ideas from Alex Tsun, Rachel Lin, Hunter Schafer & myself ☺

third quiz out 2 weeks from today
final on Dec 13.

Agenda

- Maximum Likelihood Estimation
- Continuous random variables
- Properties of estimators

Parameter Estimation – Workflow

Ben(θ)
Poisson(θ)
exp(θ)

Distribution $\mathbb{P}(x|\theta)$ → Independent samples $x_1, \ldots, x_n$ from $\mathbb{P}(x|\theta)$ → **Algorithm** → $\hat{\theta}$

Parameter estimate

θ = unknown parameter

Maximum Likelihood Estimation (MLE). Given data $x_1, \ldots, x_n$, find $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ ("the MLE") such that $L(x_1, \ldots, x_n | \hat{\theta})$ is maximized!

Likelihood of Different Observations

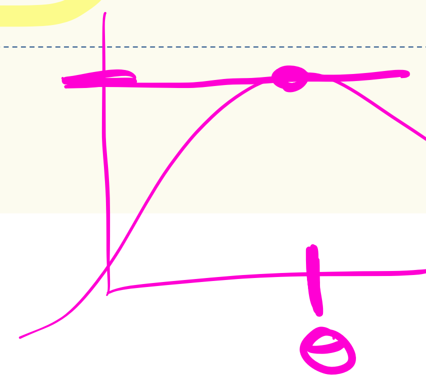
(Discrete case)

Definition. The **likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \mathbb{P}(x_i; \theta)$$

Maximum Likelihood Estimation (MLE). Given data $x_1, \ldots, x_n$, find $\hat{\theta}$ ("the MLE") of model such that $L(x_1, \ldots, x_n | \hat{\theta})$ is maximized!

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \, \mathcal{L}(x_1, \ldots, x_n | \theta)$$



$$\mathcal{L}(x_1, \ldots x_n | \theta) = P(x_1; \theta) P(x_2; \theta) \ldots \quad P(x_n; \theta)$$

H, T

Ber(θ)

$\boxed{HHTH \ldots T}$

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$L(x_1, \ldots, x_n | \theta) = \theta^{n_H} (1 - \theta)^{n_T}$$

$$\frac{\partial}{\partial \theta} L(x_1, \ldots, x_n | \theta) = ???$$

While this derivative is not hard to compute, we make our lives easier by observing that we are always taking a derivative of a product…. And logs turn products into sums…

Log-Likelihood

Save some work using **log-likelihood** instead of the likelihood directly.

Definition. The **log-likelihood** of independent observations $x_1, \dots, x_n$ is

$$\mathcal{LL}(x_1, \dots, x_n|\theta) = \ln \mathcal{L}(x_1, \dots, x_n|\theta)$$

$$= \ln \prod_{i=1}^{n} \mathbb{P}(x_i; \theta) = \sum_{i=1}^{n} \ln \mathbb{P}(x_i; \theta)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(x_1, \dots, x_n|\theta) = \underset{\theta}{\operatorname{argmax}}\, \mathcal{LL}(x_1, \dots, x_n|\theta)$$

log monotone ↑ fn.

$a > b$

$\ln a > \ln b$

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$\ln(\theta^{n_H}) + \ln((1-\theta)^{n_T})$

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \theta^{n_H}(1-\theta)^{n_T}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n|\theta) = n_H \ln\theta + n_T \ln(1-\theta)$$

$$\boxed{\frac{d}{d\theta} LL = \frac{n_H}{\theta} + \frac{n_T}{1-\theta}(-1)} = 0$$

Useful log properties.

$$\log(ab) = \log(a) + \log(b)$$
$$\log(a/b) = \log(a) - \log(b)$$
$$\log(a^b) = b\log(a)$$

$$\frac{n_H}{\hat\theta} - \frac{n_T}{1-\hat\theta} = 0$$

$$\frac{d}{d\theta} LL(x \ldots) = 0$$

$$\frac{n_H}{\hat\theta} = \frac{n_T}{1-\hat\theta} \Rightarrow (1-\hat\theta)n_H = \hat\theta n_T$$

$n_H = \hat{\theta}(\underbrace{n_H + n_T}_{n})$

$\Rightarrow \hat{\theta} = \frac{n_H}{n}$

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \theta^{n_H} (1-\theta)^{n_T}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n | \theta) = n_H \ln \theta + n_T \ln(1-\theta)$$

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n | \theta) = n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1-\theta}$$

Solve $n_H \cdot \frac{1}{\hat{\theta}} - n_T \cdot \frac{1}{1-\hat{\theta}} = 0$

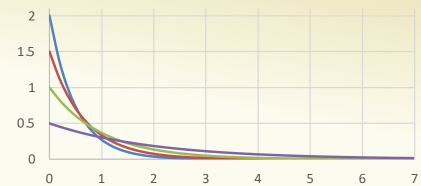
$$\boxed{\hat{\theta} = \frac{n_H}{n}}$$

Brain Break

Agenda

- Maximum Likelihood Estimation (Recap + LogLikelihood)
- Continuous random variables
- Properties of estimators

The Continuous Case



0.11, 0.35, ...

Given n samples $x_1, \dots, x_n$ from an exponential distribution with unknown parameter θ

Definition. The **likelihood** of independent observations $x_1, \dots., x_n$ is

$$\mathcal{L}(x_1, \dots., x_n | \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Density function! (Why?)

Why density?

- Density ≠ probability, but:
  - For maximizing likelihood, we really only care about relative likelihoods, and density captures that
  - Estimates probability of seeing samples close to $x_1, \dots, x_n$

$$\Pr\left(X_i \in [x_i, x_i + dx]\right) \approx f(x_i)\, dx$$

MLE for exponential distribution

$X \sim \exp(\lambda)$ $f(x) = \lambda e^{-\lambda x}$ $x \geq 0$

Given n samples $x_1, \ldots, x_n$ from an Exponential distribution with unknown parameter θ

The **likelihood** function of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

Find the MLE $\hat{\theta}$

$$\mathcal{L} = \theta^n e^{-\theta x_1 - \theta x_2 \cdots - \theta x_n}$$

$$\ln(\theta^n) + \ln(e^{-\theta \Sigma x_i})$$

$$LL(x_1, \dots, x_n \mid \theta) = n \ln \theta - \theta \sum_{i=1}^{n} x_i \underbrace{\cancel{\ln e}}_{1}$$

$\log(ab) = \log(a) + \log(b)$

$\log(a/b) = \log(a) - \log(b)$

$\log(a^b) = b\log(a)$

$\boxed{\ln(e) = 1}$

$$\frac{d}{d\theta} LL = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$

$$\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

$$\frac{n}{\hat{\theta}} = \sum_{i=1}^{n} x_i \Rightarrow \hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i}$$

λ of exp distn

note $E(\hat{\theta}) \neq \lambda$

General Recipe

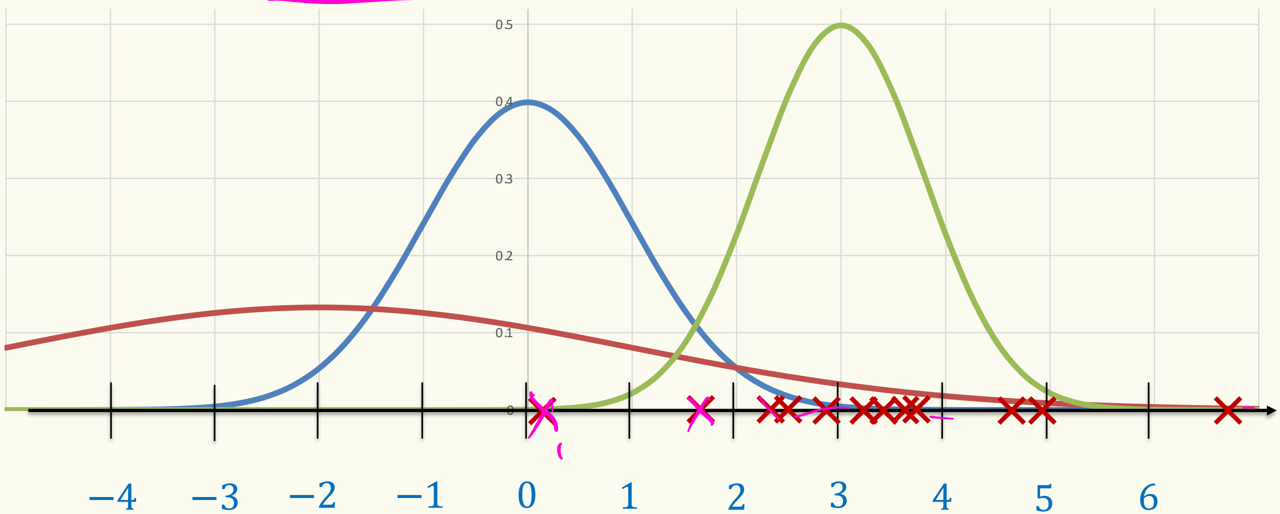
1. **Input** Given n iid samples $x_1, \ldots, x_n$ from parametric model with parameters θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \ldots, x_n|\theta)$.
   - For discrete $\mathcal{L}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \Pr(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \ldots, x_n|\theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial\theta} \ln \mathcal{L}(x_1, \ldots, x_n|\theta)$
5. **Solve for** $\hat{\theta}$ by setting derivative to 0 and solving for max.

Generally, you need to do a second derivative test to verify it is a maximum, but we won't ask you to do that in CSE 312.

unknown

Next: n samples $x_1, \ldots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, \sigma^2)$.

Most likely μ and σ^2?



$$\frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}$$

$\ln(e^{\text{some}}) = \boxed{\text{somet}}$

$L(x_1 \dots x_n | \theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i ; \theta_1, \theta_2)$

Two-parameter optimization

$\log(ab) = \log(a) + \log(b)$

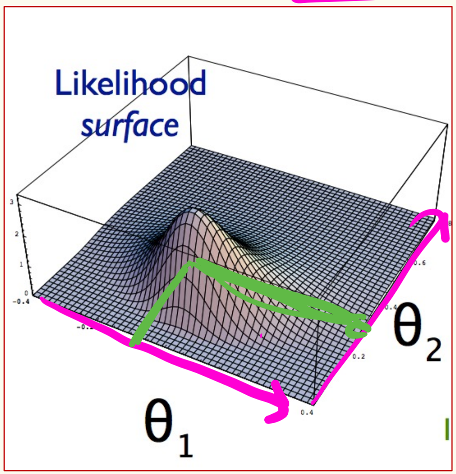
$\log(a/b) = \log(a) - \log(b)$

$\log(a^b) = b\log(a)$

$\ln(e) = 1$

Normal outcomes $x_1, \dots, x_n$

Goal: estimate $\boxed{\theta_1 = \mu}$ = expectation and $\boxed{\theta_2 = \sigma^2}$ = variance



$$L(x_1, \dots, x_n | \theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}$$

$$\ln L(x_1, \dots, x_n | \theta_1, \theta_2) =$$

$$n \ln\left(\frac{1}{\sqrt{2\pi\theta_2}}\right) - \sum_{i=1}^{n} \frac{(x_i-\theta_1)^2}{2\theta_2}$$

$$-n \ln\left(\sqrt{2\pi\theta_2}\right)$$

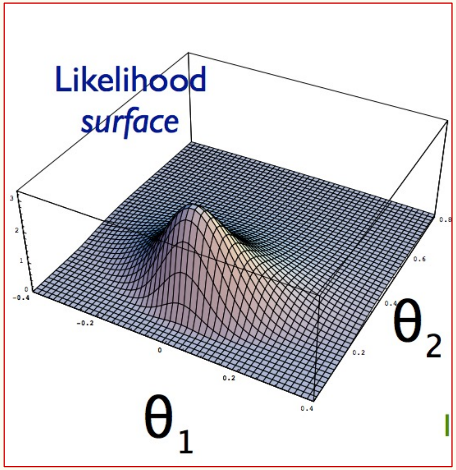
$(2\pi\theta_2)^{1/2}$

$$-\frac{n}{2}\ln(2\pi\theta_2) - \sum_{i=1} \frac{(x_i-\theta_1)}{2\theta_2}$$

Two-parameter optimization

Normal outcomes $x_1, \dots, x_n$

Goal: estimate $\theta_1 = \mu$ = expectation and $\theta_2 = \sigma^2$ = variance



$$L(x_1, \dots, x_n | \theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}$$

$$\ln L(x_1, \dots, x_n | \theta_1, \theta_2) =$$

$$= -n\frac{\ln(2\pi\,\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i-\theta_1)^2}{2\theta_2}$$

Two-parameter estimation

$$\ln L(x_1, \ldots., x_n | \theta_1, \theta_2) = -n\frac{\ln(2\pi\,\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

We need to find a solution $\hat{\theta}_1, \hat{\theta}_2$ to

$$\boxed{\begin{aligned} \frac{\partial}{\partial\theta_1} \ln L(x_1, \ldots., x_n | \theta_1, \theta_2) &= 0 \\ \frac{\partial}{\partial\theta_2} \ln L(x_1, \ldots., x_n | \theta_1, \theta_2) &= 0 \end{aligned}}$$

$\hat{\theta}_1, \hat{\theta}_2$

MLE for Expectation

$$\ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = -n\frac{\ln(2\pi\theta_2)}{2} - \sum_{i=1}^{n} \frac{\boxed{(x_i - \theta_1)^2}}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = -\frac{1}{2\theta_2} \sum_{i=1}^{n} 2(x_i - \theta_1)(-1)$$

$$\frac{1}{\hat{\theta}_2} \sum_{i=1}^{n} (x_i - \hat{\theta}_1) = 0 \qquad / \hat{\theta}_2$$

$$\sum_{i=1}^{n} (x_i - \hat{\theta}_1) = 0$$

$$\sum_{i=1}^{n} x_i - n\hat{\theta}_1 = 0$$

$$\boxed{\hat{\theta}_1 = \frac{\sum_{i=1}^{n} x_i}{n}}$$

MLE for Expectation

$$\ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = -n\frac{\ln(2\pi\,\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, \ldots, x_n | \theta_1, \theta_2) = \frac{1}{\theta_2} \sum_{i}^{n} (x_i - \theta_1) = 0$$

$$\hat{\theta}_1 = \frac{\sum_i^n x_i}{n}$$

In other words, MLE of expectation is the *sample mean* of the data, regardless of θ_2

What about the variance?

$\ln(2\pi\theta_2) = \ln(2\pi) + \ln(\theta_2)$

MLE for Variance

$$\ln L(x_1, \dots, x_n | \hat{\theta}_1, \theta_2) = -n \frac{\ln(2\pi\,\theta_2)}{2} - \sum_{i=1}^{n} \frac{(x_i - \hat{\theta}_1)^2}{2\theta_2}$$

$$= \cancel{-n \frac{\ln 2\pi}{2}} - n \frac{\ln \theta_2}{2} - \frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2$$

$\frac{1}{\theta_2}$ still.

$= \frac{1}{\theta_2^2}$

$$\frac{\partial}{\partial \theta_2} \ln L(x_1, \dots, x_n | \theta_1, \hat{\theta}_1) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2 = 0$$

$$\boxed{\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2}$$

In other words, MLE of variance is what's called the *population variance* of the data.

Likelihood – Continuous Case

Definition. The **likelihood** of independent observations $x_1, \dots, x_n$ is

$$L(x_1, \dots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta)$$

Normal outcomes $x_1, \dots, x_n$

$\hat{\theta}_1$

$$\hat{\theta}_\mu = \frac{\sum_i^n x_i}{n}$$

unbiased estimator

MLE estimator for **expectation**

$\hat{\theta}_2$

$$\hat{\theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_\mu\right)^2$$

MLE estimator for **variance**

$$\mathop{E}_{X_1 \dots X_n}\left(\frac{\sum X_i}{n}\right) = \mu$$

Agenda

- Maximum Likelihood Estimation (Recap + LogLikelihood)
- Continuous random variables
- Properties of estimators

When is an estimator good?

Distribution $p(x|\theta)$ → samples $X_1, \dots, X_n$ from $p(x|\theta)$ → **Algorithm** → $\hat{\theta}_n$

Parameter estimate "The model"

θ = unknown parameter

$X_1 \dots X_n$

Definition. An estimator of parameter θ is an **unbiased estimator**

$$\boxed{\mathbb{E}(\hat{\theta}_n)} = \theta.$$

$X_1, X_2, \dots X_n$

Ber(p)

Example – Coin Flips

Recall: $\hat{\theta}_{\boldsymbol{\mu}} = \frac{n_H}{n}$

Coin-flip outcomes $x_1, \dots, x_n$, with n_H heads, n_T tails

Fact. $\hat{\theta}_{\boldsymbol{\mu}}$ is unbiased

i.e., $\mathbb{E}(\hat{\theta}_{\boldsymbol{\mu}}) = p$, where p is the probability that the coin turns out heads.

Consistent Estimators & MLE

Parameter estimate

Distribution $\mathbb{P}(x|\theta)$ → samples $X_1, \ldots, X_n$ from $\mathbb{P}(x|\theta)$ → **Algorithm** → $\hat{\theta}_n$

θ = unknown parameter

Definition. An estimator is **unbiased** if $\mathbb{E}(\hat{\theta}_n) = \theta$ for all $n \geq 1$.

Definition. An estimator is **consistent** if $\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta$.

Theorem. MLE estimators are consistent.

(But not necessarily unbiased)

Example – Consistency

Normal outcomes $X_1, \dots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

$$\widehat{\Theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_\mu\right)^2$$

MLE – Biased!

$\widehat{\Theta}_{\sigma^2}$ converges to σ^2, as $n \to \infty$.

$\widehat{\Theta}_{\sigma^2}$ is “consistent”

Why is the estimator consistent, but biased?

linearity

$$\mathbb{E}(\widehat{\Theta}_{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \widehat{\Theta}_\mu\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2 - \frac{2}{n}X_i\sum_{j=1}^{n}X_j + \frac{1}{n^2}\sum_{j=1}^{n}X_j\sum_{k=1}^{n}X_k\right]$$

…

Why is the estimator consistent, but biased?

linearity

$$\mathbb{E}(\widehat{\Theta}_{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(X_i - \widehat{\Theta}_1\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2\right]$$

…

$$= \left(1 - \frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2$$

Why is the estimator consistent, but biased?

linearity

$$\mathbb{E}(\widehat{\Theta}_{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \widehat{\Theta}_1\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right]$$

…

$$= \left(1 - \frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2 \;\to \sigma^2 \text{ for } n \to \infty$$

Therefore: $\frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \widehat{\Theta}_1\right)^2\right] = \frac{n}{n-1}\mathbb{E}(\widehat{\Theta}_{\sigma^2}) = \sigma^2$

Bessel's correction

Example – Consistency

Normal outcomes $X_1, \dots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

$$\widehat{\Theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

MLE – Biased!

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Sample variance – Unbiased!

$\widehat{\Theta}_{\sigma^2}$ converges to σ^2, as $n \to \infty$.

$\widehat{\Theta}_{\sigma^2}$ is "consistent"