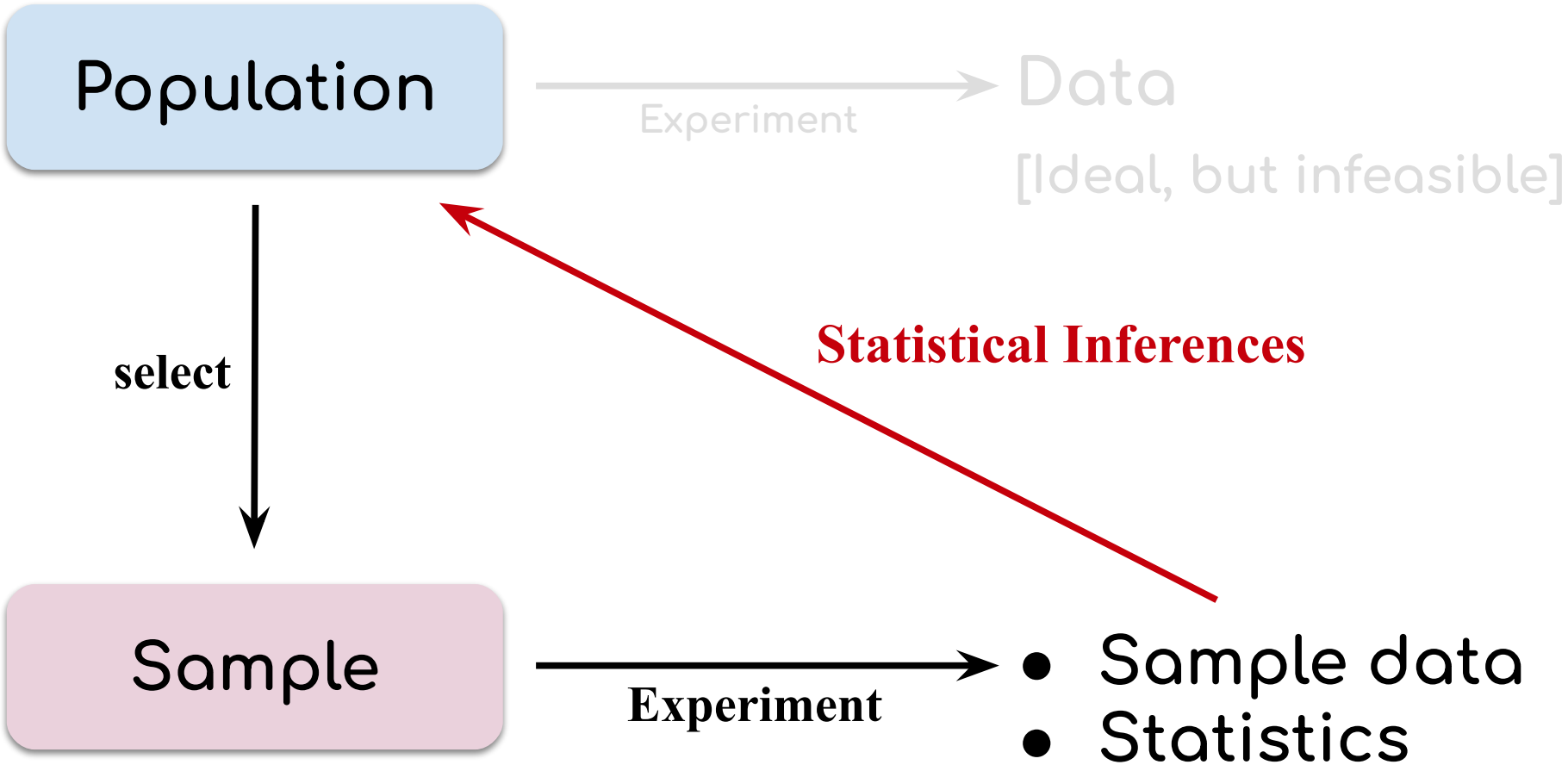
통계분석
Statistical Analysis

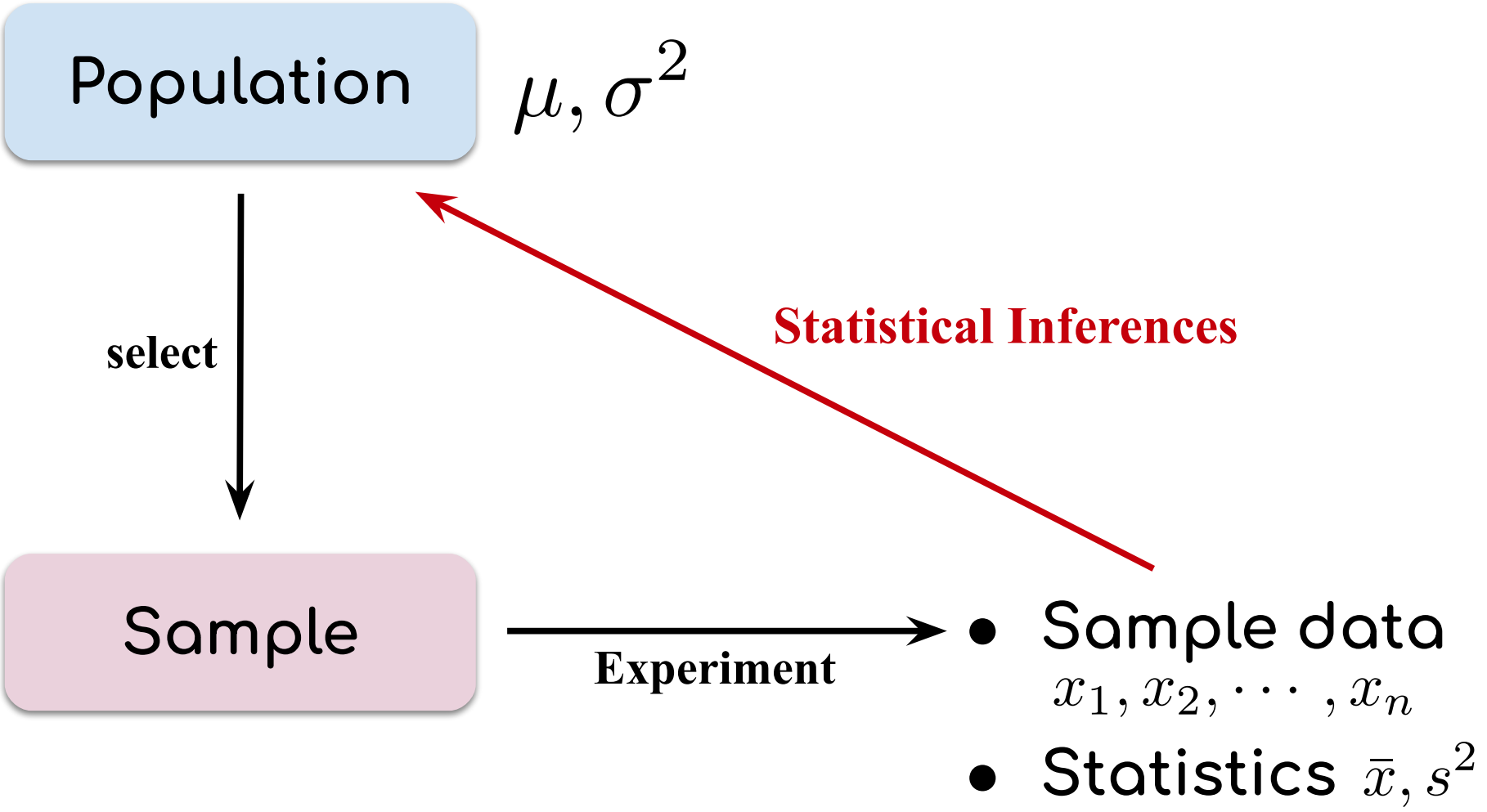
Statistical Inferences

Recall: Statistical Inferences



Statistical Inference: Techniques for generalizing from a sample to a population

Recall: Statistical Inferences



Statistics Inferences

1. **Estimation**

   - **<u>Point Estimation</u>: single value estimate from sample data**

$$\text{mean } \mu, \text{variance } \sigma^2$$

   - **<u>Interval Estimation</u>: [a,b] interval estimate**

Confidence interval for population mean

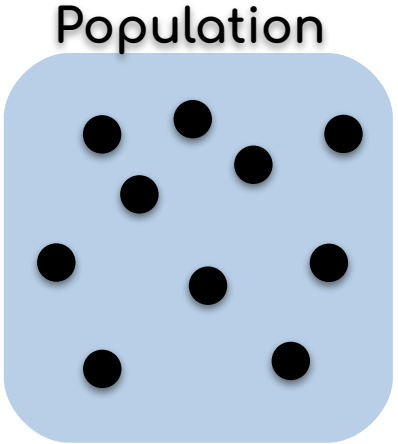
$$a < \mu < b$$

2. **Hypotheses Testing**

- **Hypotheses on (Population) Parameters**
  (ex) Population mean is a.
- **Testing is to verify whether the hypotheses are true or not.**

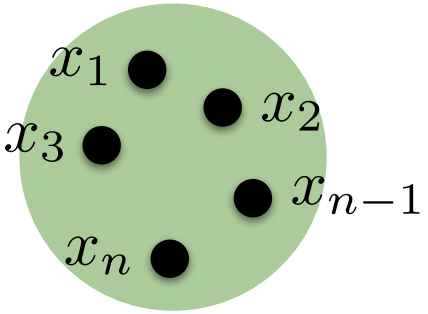
Point Estimation

Point Estimation



Population

Parameters (모수) θ



Sample (표본)

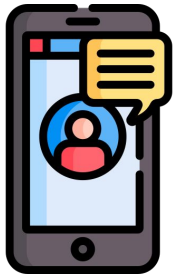
$\{x_1, x_2, \cdots, x_{n-1}, x_n\}$

Some statistic $= f(x_1, x_2, \cdots, x_n)$

Point Estimation **: the sample statistic regarded as the parameter that we want to know**

Point estimation : $f = \theta$

Point Estimation : Example



An IT company manufactures 100,000 smartphones this year.
We want to estimate the average of lifetime of those phones.
We cannot test all of them. Instead, we choose several samples from them randomly, and we do a lifetime test on the samples.

For example, we choose five phones.

$$x_1 = 2.5, x_2 = 2.2, x_3 = 3.1, x_4 = 1.7, x_5 = 2.5$$

$$\text{sample mean} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 2.4$$

- Regard this number as the population mean → Point estimation
- The sample mean can be one possible example of point estimates for the true population mean
- Even though the sample mean is quite a reasonable candidate for point estimate, it is noticed that it is not the only one.

Point Estimator and Point Estimate

- **Definition: Point Estimator**

$X_1, X_2, \cdots, X_n =$ Independent random variables from the population whose parameter is θ

$f(X_1, X_2, \cdots, X_n) =$ a statistic used to estimate θ

$$\hat{\theta} = f(X_1, X_2, \cdots, X_n) = \text{Point Estimator of } \theta$$

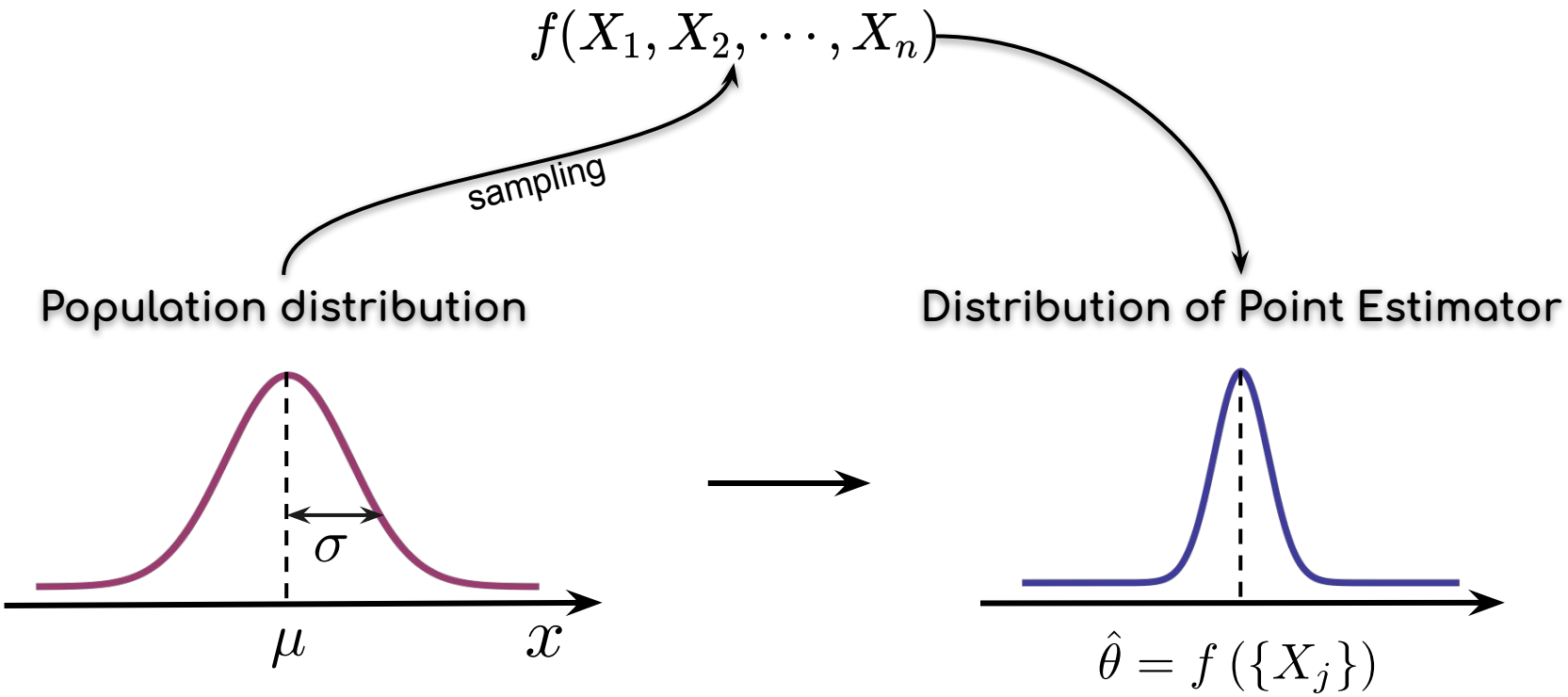
- **Definition: Point Estimates**

For a specific sample whose data are $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$

$$f(x_1, x_2, \cdots, x_n) = \text{Point Estimate of } \theta$$

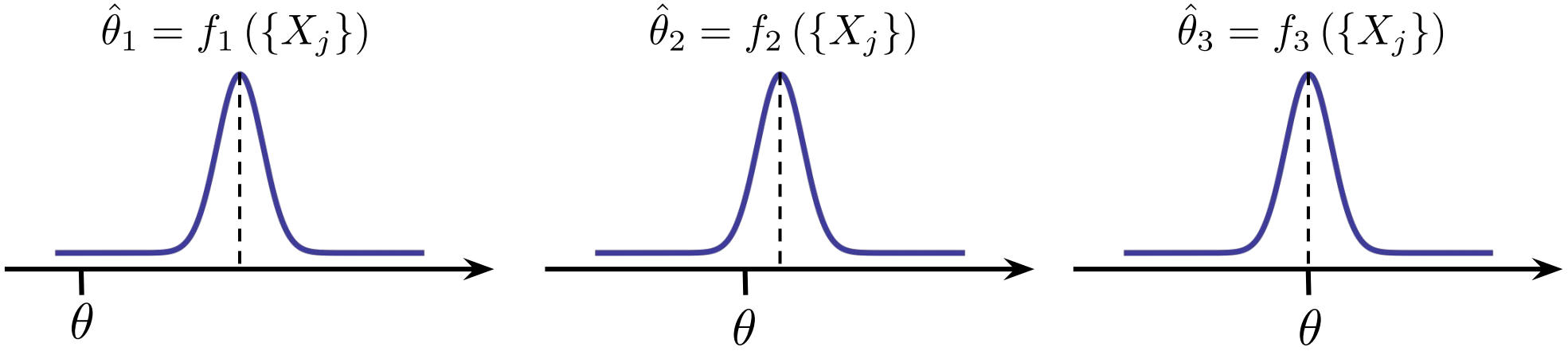
- 모수(parameters)를 추정하기 위하여 사용한 표본통계량 함수를 Point Estimator라고 한다.
- 특정 표본 자료를 표본통계량 함수에 대입하여 계산한 수를 Point Estimate라고 한다.
  Point Estimate는 표본에 따라 다른 값을 가진다.

Distribution of Point Estimator (Statistic)

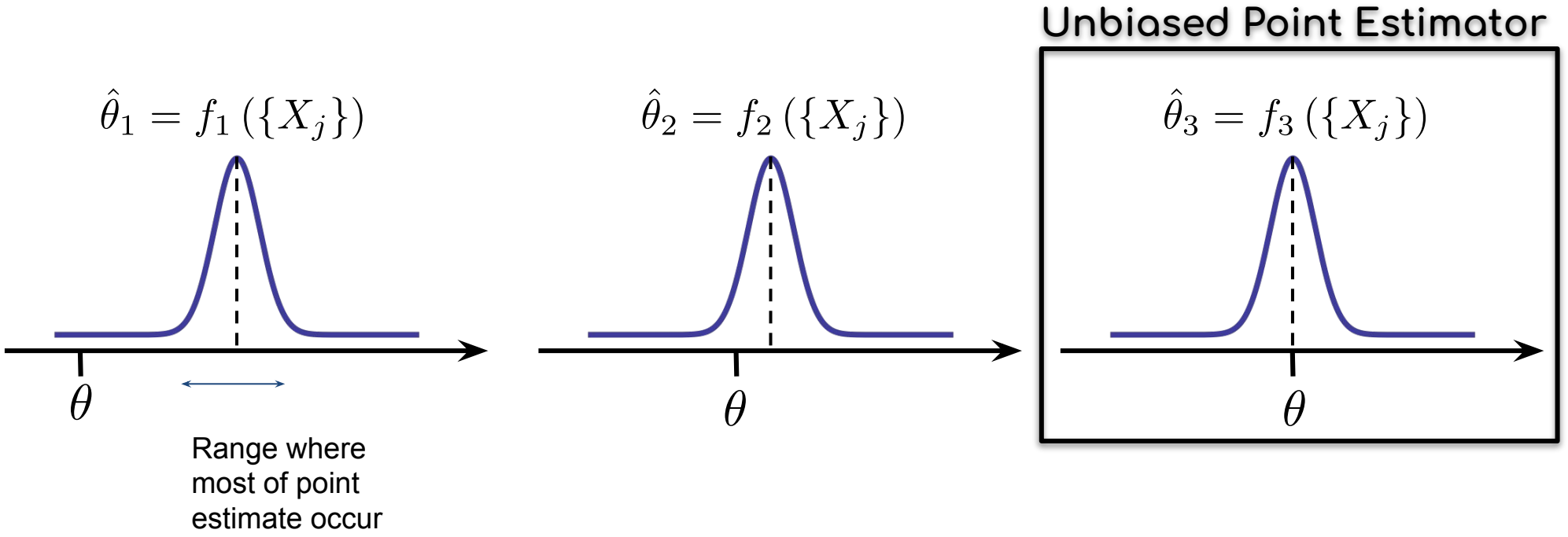


- Point estimator, which is one of statistics, follows a probability distribution.
- The value of the point estimator varies sample by sample.
- Some point estimate may be close to the population parameter, while another point estimate may be far from the population parameter.

Condition I for Better Point Estimator



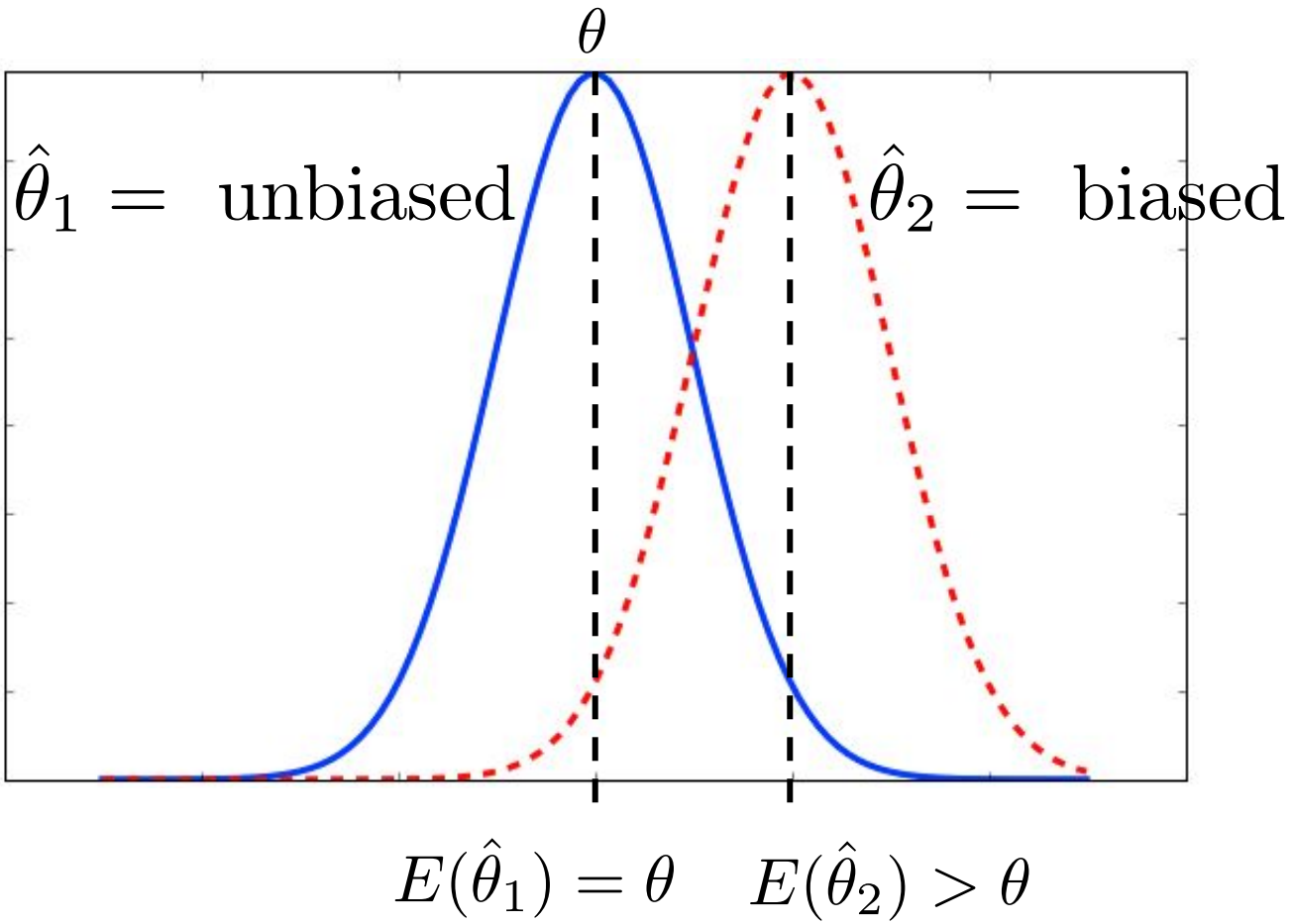
Unbiased Point Estimator



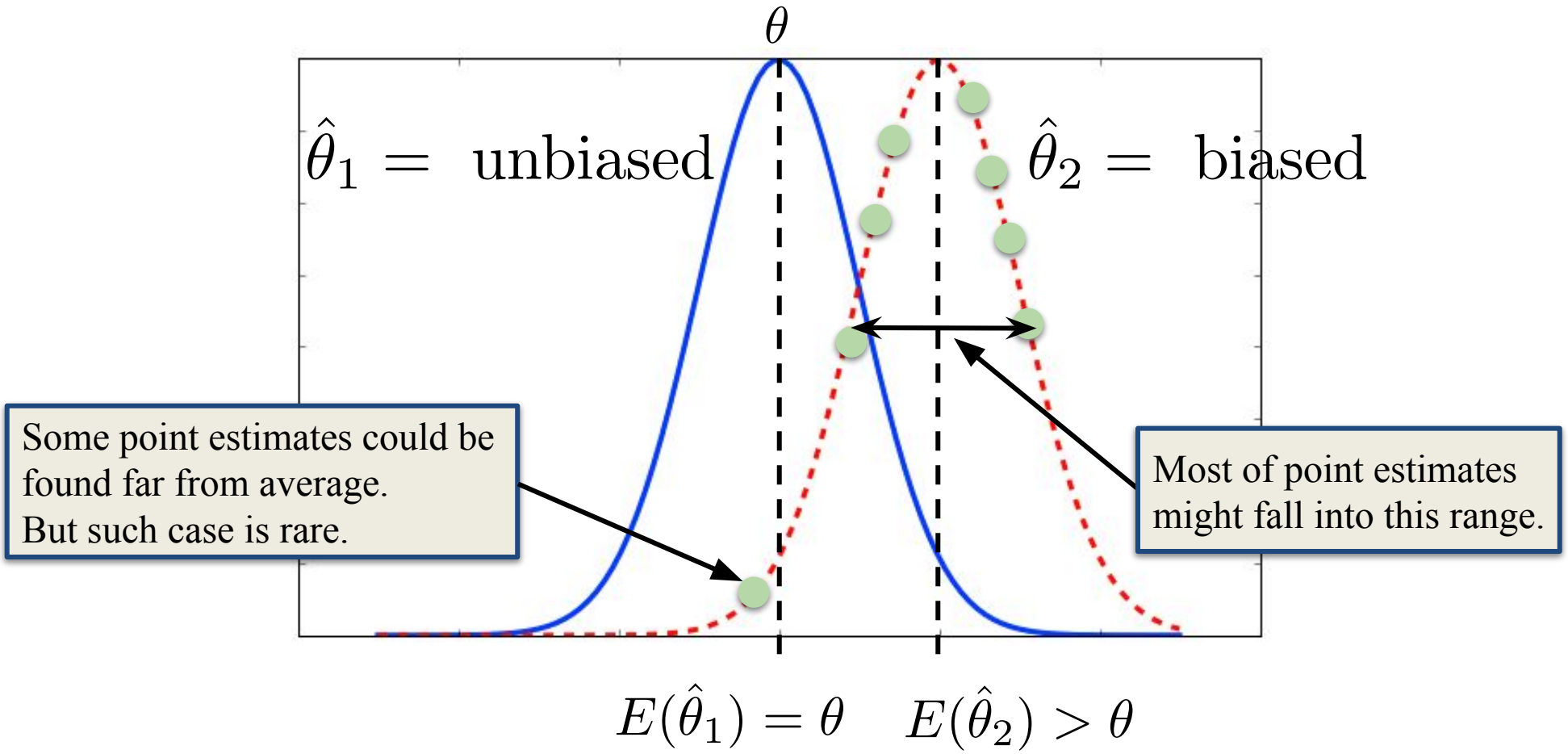
$$E(\hat{\theta}) = \theta : \text{Unbiased estimator}$$

$$E(\hat{\theta}) \neq \theta : \text{Biased estimator}$$

Distributions of Estimators : Unbiased vs Biased



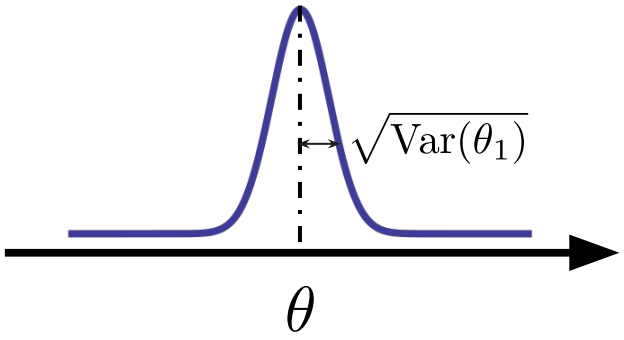
Distributions of Estimators : Unbiased vs Biased



Biased point estimator is likely not to give a good point estimate close to the population value.

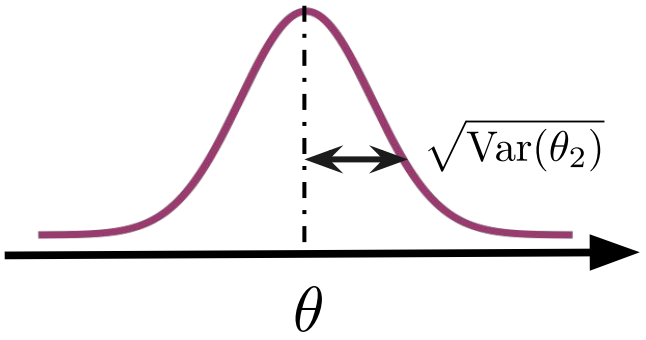
Condition II for Better Point Estimator

$$\hat{\theta}_1 = f_1\left(\{X_j\}\right)$$



$$E(\hat{\theta}_1) = \theta$$

$$\hat{\theta}_2 = f_2\left(\{X_j\}\right)$$



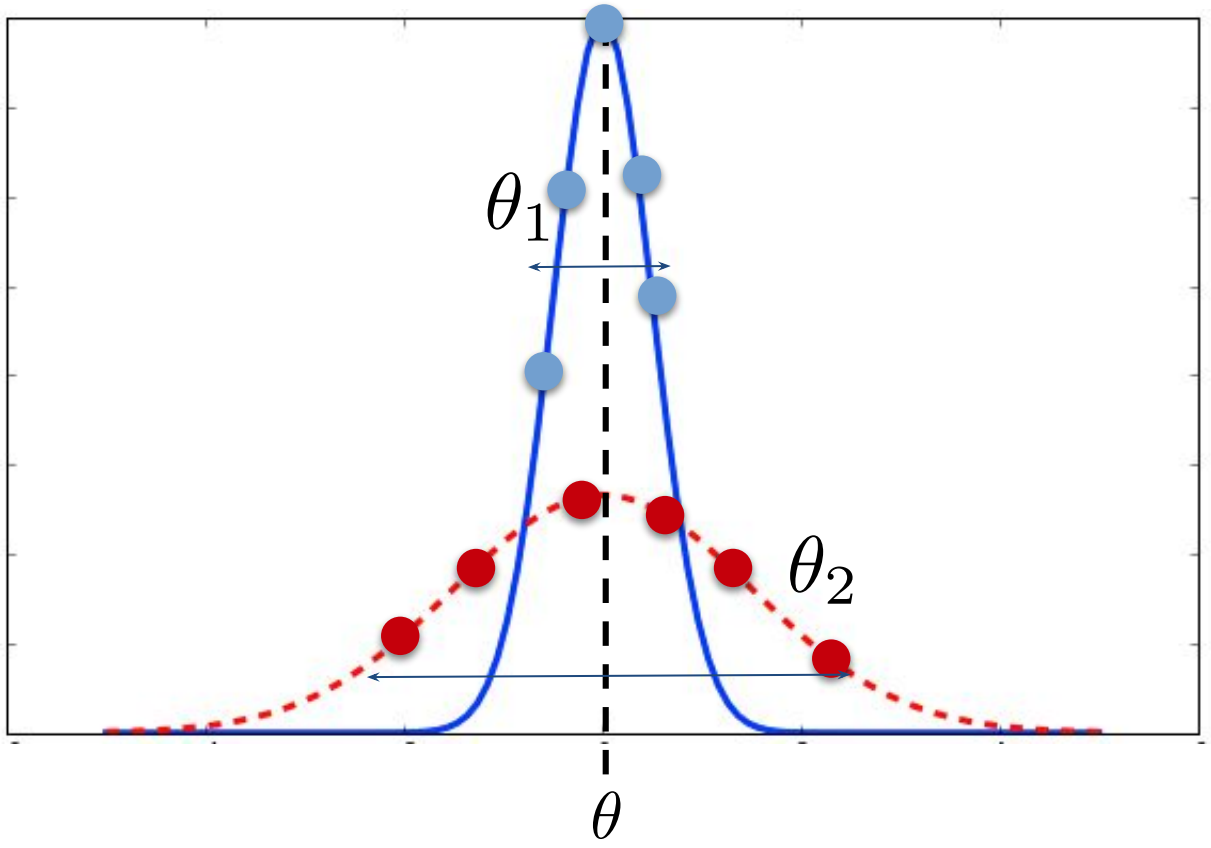
$$E(\hat{\theta}_2) = \theta$$

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$

Which unbiased estimator is better for the point estimator?

Minimum Variance Unbiased Estimator

$$E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta \quad \mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$



Unbiased estimator with smaller variance is more likely to give the estimate close to the population value than unbiased estimator with larger variance.

Unbiased Point Estimator of Population Mean

$X_1, X_2, \cdots, X_n$: Random sample from a distribution with mean μ and variance σ^2

$$\text{Sample Mean} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \begin{cases} E(\bar{X}) = \mu_{\bar{X}} = \mu \\ V(\bar{X}) = \sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n} \\ \sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} \end{cases}$$

Population paramter to estimate $= \theta = \mu$

$$\text{Point estimator} = \hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i : \text{Sample Mean}$$

- $E(\hat{\theta}) = E(\bar{X}) = \mu$: unbiased
- $\text{Var}(\hat{\theta}) = \dfrac{\sigma^2}{n}$: As $n \to \infty, \text{Var}(\hat{\theta}) \to 0$

Minimum variance unbiased point estimator for population mean = sample mean with very large n

Unbiased Point Estimator of Population Variance

$X_1, X_2, \cdots, X_n =$ Independent random variables
from the population whose parameter is θ

Population parameter : $\theta = \sigma^2$

$$\hat{\theta}_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2? \qquad \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2?$$

Unbiased Point Estimator of Population Variance

$X_1, X_2, \cdots, X_n =$ Independent random variables
from the population whose parameter is θ

Population parameter : $\theta = \sigma^2$

$$\hat{\theta}_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2? \qquad \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2?$$

Unbiased estimator **Biased estimator**

$$E(\hat{\theta}_1) = \sigma^2 \qquad E(\hat{\theta}_2) = \frac{n-1}{n}\sigma^2 < \sigma^2$$

- Unbiased point estimator is regarded as Sample variance, because it is more suitable to estimation of population variance.
- The biased estimator defined above underestimates the population variance.

Unbiased Point Estimator of Population Variance

$X_1, X_2, \cdots, X_n =$ Independent random variables
from the population whose parameter is θ

Population parameter : $\theta = \sigma^2$

$$\hat{\theta} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- $E(\hat{\theta}) = E(S^2) = \sigma^2$
- $\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(S^2) = \dfrac{2\sigma^4}{n-1}$: As $n \to \infty$, $\mathrm{Var}(S^2) \to 0$

- Minimum variance unbiased point estimator for population variance = sample variance with very large n

Point Estimator of Population /Mean Variance

$$\text{Point estimator} = \hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ : \ \text{Sample Mean}$$

$$\hat{\theta} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad : \text{Sample Variance}$$

- These two estimators are unbiased, and minimum variances of the two estimators are achieved by increasing the sample size n.

- Problems of Point Estimation : Hard to describe quantified error that your estimation is not close to the true parameter

- ***Interval estimation*** : you can estimate a range where the population parameter is located and you can also provide an error that your estimation is wrong.