Chapter 7: Sampling Distributions and Point Estimation of Parameters

Topics:

- General concepts of estimating the parameters of a population or a probability distribution
- Understand the central limit theorem
- Explain important properties of point estimators, including bias, variance, and mean square error

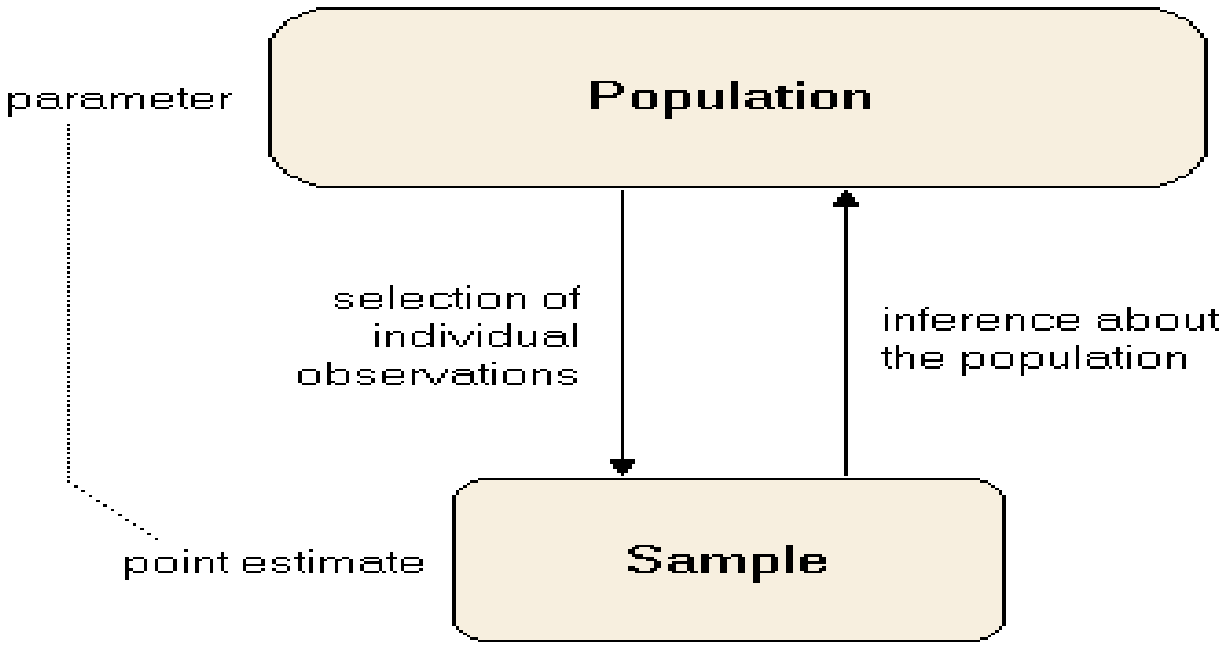
Overview

- Identify a population of interest
  —-for example, UNM freshmen female students' weight, height or entrance GPA.
- Population parameters
  —-unknown quantities of the population that are of interest, say, population mean μ and population variance σ^2 etc.
- Random sample
  —-Select a random or representative sample from the population.
  —-A sample consists random variables $Y_1, \cdots, Y_n$, that follows a specified distribution, say $N(\mu, \sigma^2)$
- Statistic: a function of radom variables $Y_1, \ldots, Y_n$, which does not depend on any unknown parameters
- Observed sample: $y_1, y_2, \cdots, y_n$ are observed sample values after data collection

- We cannot see much of the population
  —-but would like to know what is typical in the population
  — The only information we have is that in the sample.

Goal: want to use the sample information to make inferences about the population and its parameters.

- Statistical inference is concerned with making decisions about a population based on the information contained in a random sample from that population.



Point estimation

Suppose our goal is to obtain a point estimate of a population parameter, i.e. mean, variance, based a sample $x_1, \ldots, x_n$.

- Before we collected the data, we consider each observation as a random variable, i.e. $X_1, \ldots, X_n$.
- We assume $X_1, \ldots, X_n$ are mutually independent random variables.

Point estimator: a point estimator is a function of $X_1, \ldots, X_n$.
Point estimate: a point estimate is a single numerical value of the point estimator based on an observed sample.

- Population mean: μ
- Sample mean: $\bar{Y} = \sum_{i=1}^{n} Y_i/n$
- Estimate of sample mean: the value of $\bar{Y}$ computed from data $\bar{y} = \sum_{i=1}^{n} y_i/n$
- Population variance: σ^2
- Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- Estimate of sample variance: the value of S^2 computed from data $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Population standard deviation: σ
- Sample standard deviation (Standard error): S
- Estimate of standard error: s, the value of S computed from data

Table : Commonly seen parameters, statistics and estimates:

Parameters Describe a popn	Statistic Describe a random sample	Estimate Describe an observed sample
μ	$\bar{Y}$	$\bar{y}$
σ^2	S^2	s^2
σ	S	s

Example

Table 6.5 contains a second example of multivariate data taken from an article on the quality of different young red wines in the Journal of the Science of Food and Agriculture (1974, Vol. 25) by T.C. Somers and M.E. Evans. The authors reported quality along with several other descriptive variables. We are interested in quality and PH values for a sample of their wines.

```
winequality <- c(19.2, 18.3, 17.1, 15.2, 14.0, 13.8, 12.8,
 17.3, 16.3, 16.0, 15.7, 15.3, 14.3, 14.0,13.8, 12.5, 11.5,
 14.2,17.3,15.8)
PH<-c(3.85,3.75,3.88,3.66,3.47,3.75,3.92,3.97,3.76,3.98,
3.75,3.77,3.76,3.76,3.90,3.80,3.65,3.60,3.86,3.93)
```

- Give an estimate for the mean of wine quality rate (μ).
- Give an estimate for the variance of wine quality rate (σ^2).
- Give an estimate for the correlation of wine quality and PH.

Recall that Correlation between two sample data $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$:

$$r_{xy} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

measure linear relationship between x and y.

R codes to find the answers:

```
mean(winequality)
[1] 15.22

 var(winequality)
[1] 3.992211

cor(winequality,PH)
[1] 0.3492413
```

- Give an estimate for the mean of wine quality rate (μ).
  From R output, the estimate for the mean of wine quality rate (μ) is $\bar{x} = 15.22$
- Give an estimate for the variance of wine quality rate (σ^2).
  From R output, the estimate for the variance of wine quality rate (σ^2) is $s^2 = 3.99$
- Give an estimate for the correlation of wine quality and PH.
  From R output, the estimate for the correlation of wine quality and PH is 0.3492413.

Sampling distribution

Sampling distribution: probability distribution of a given statistic based on a random sample

—-Statistic is also a r.v.

—-Sampling distribution is in contrast to the population distribution

Want to know the sampling distribution of $\bar{X}$

- standard error (SE): the standard deviation of the sampling distribution of a statistic
- Standard error of the mean (SEM): is the standard deviation of the sample-mean's estimator

If $X_1, \ldots, X_n$ are observations of a random sample of size n from normal distributions $N(\mu, \sigma^2)$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean of the n observations. We have

$$SE_{\bar{X}} = s/\sqrt{n}$$

where
s is the sample standard deviation (i.e., the sample-based estimate of the standard deviation of the population)
n is the size (number of observations) of the sample.

Central limit theorem (CLT)

If $X_1, \ldots, X_n$ is a random sample of size n taken from a population or a distribution with mean μ and variance σ^2 and if $\bar{X}$ is the sample mean, then for large n,

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Standardization

If $X_1, \ldots, X_n$ is a random sample of size n taken from a normal population with mean μ and variance σ^2 and if $\bar{X}$ is the sample mean, then,

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

We may standardize $\bar{X}$ by subtracting the mean and dividing by the standard deviation, which results in the variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

and

$$Z \sim N(0, 1)$$

illustration of CLT

- Consider random variables $X_i \sim$ Uniform$(0, 1)$ distribution
  —- any value in the interval $[0, 1]$ is equally likely
  —- $\mu = E(X) = 1/2$, and $\sigma^2 = Var(X) = 1/12$, so the standard deviation is $\sigma = \sqrt{1/12} = 0.289$.
- Draw a sample of size n
  —- the standard error of the mean will be $\sigma/\sqrt{n}$
  —- as n gets larger the distribution of the mean will increasingly follow a normal distribution.
  Illustration:

1. generate unifrom random sample of size n
2. calculate sample mean $\bar{y}$
3. repeat for $N = 10000$ times
4. plot those N means, compute the estimated SEM

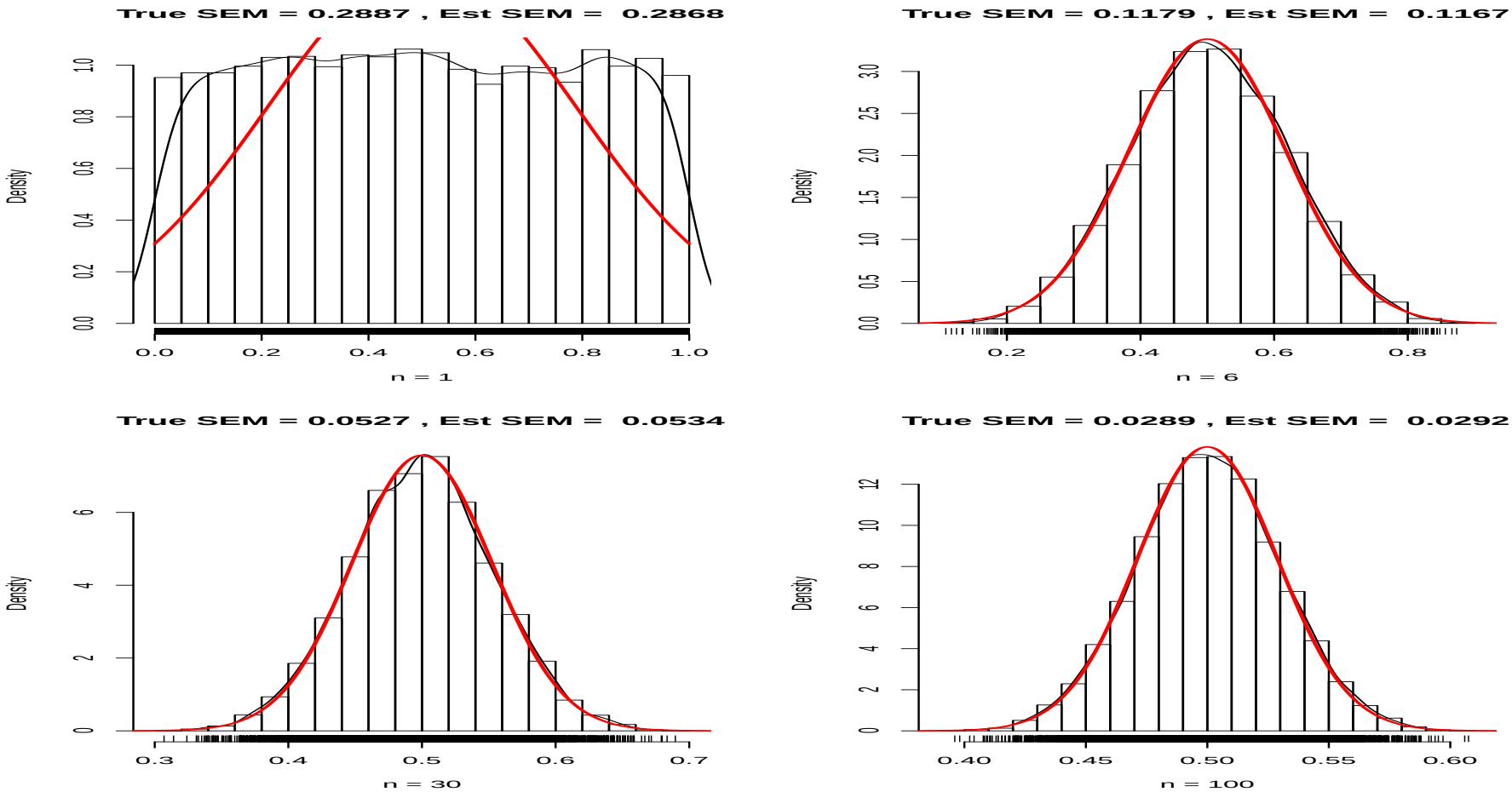


Figure : illustration of CLT, notice even with samples as small as 2 and 6 that the properties of the SEM and the distribution are as predicted

illustration of CLT

In a more extreme example, we draw samples from an Exponential(1) distribution ($\mu = 1$ and $\sigma = 1$), which is strongly skewed to the right.

$$f(x) = e^{-x}, x > 0$$

Notice that the normality promised by the CLT requires larger samples sizes, about $n \geq 30$, than for the previous Uniform(0,1) example, which required about $n \geq 6$.

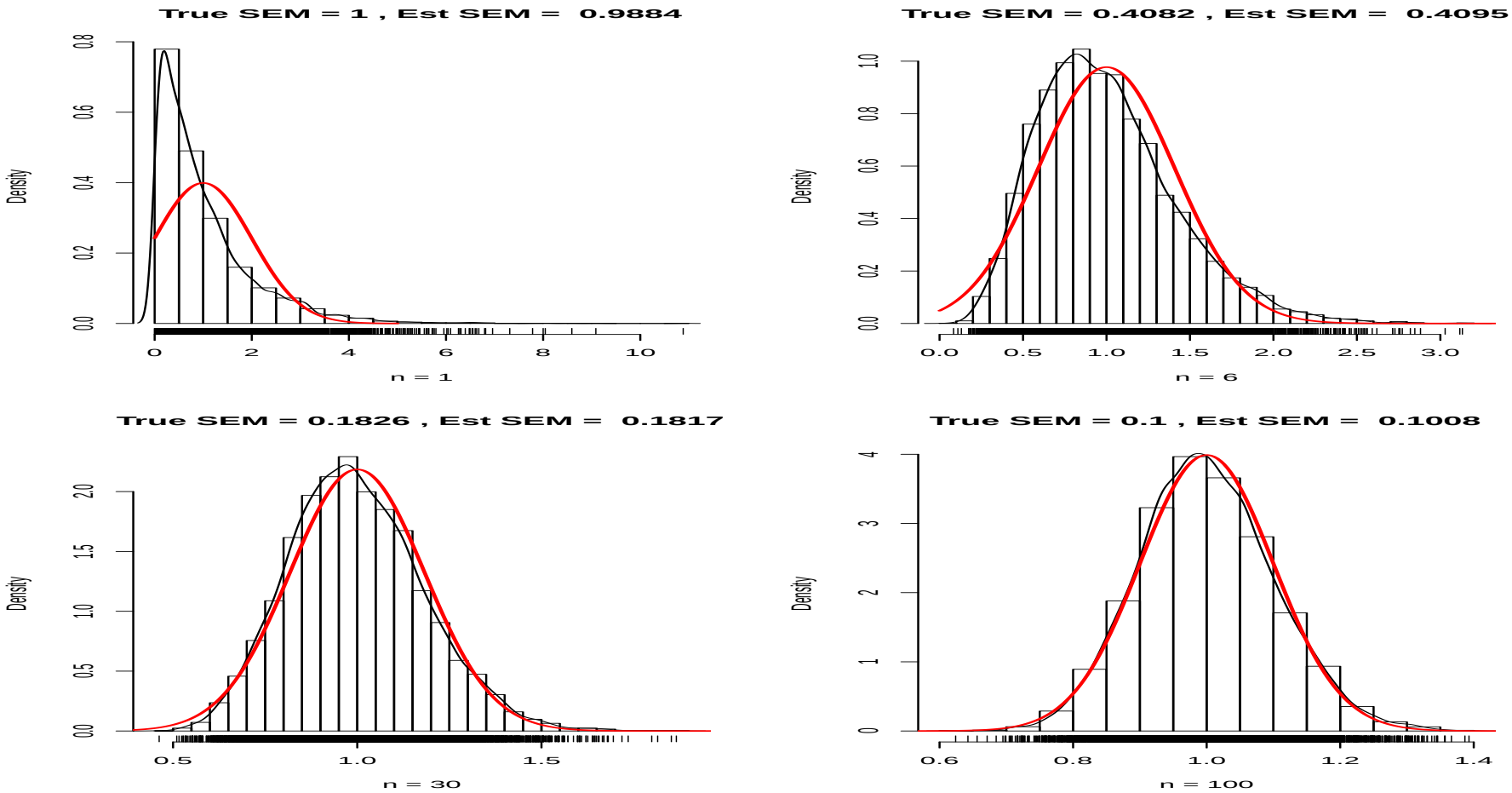


Figure : illustration of CLT, notice that the normality promised by the CLT requires larger samples sizes, about $n \geq 30$

Note that the further the population distribution is from being normal, the larger the sample size is required to be for the sampling distribution of the sample mean to be normal.

——-$n \geq 30$, normal approximation will be satisfactory regardless of the shape of the population

——$n < 30$, CLT work if the distribution of the population is not severely nonnormal.

Question: If the population distribution is normal, what's the minimum sample size for the sampling distribution of the mean to be normal?

Example:

Suppose that a r.v. X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2 & 4 \leq x \leq 6 \\ 0 & \textit{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 40$.

Solution: X has a continuous uniform distribution,

$$\mu = \frac{4+6}{2} = 5, \sigma^2 = \frac{(6-4)^2}{12} = 1/3$$

Since $n = 40$ is large, according to CLT,

$$\bar{X} \sim N(\mu, \sigma^2/n) = N(5, 1/120)$$

More on sampling distribution

- If $X_1, \ldots, X_n$ are observations of a random sample of size n from normal distributions $N(\mu, \sigma^2)$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean of the n observations. Let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance then
  - $\bar{X} \sim N(\mu, \sigma^2/n)$
  - $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$

- Two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. If $\bar{X}_1$ and $\bar{X}_2$ are the sample means of two independent random samples of size n_1 and n_2 from these two populations, then the sampling distribution of

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2).$$

  If the two populations are normal, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is exactly normal.

- If n is large, the distribution of

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

Example: The effective life of a component used in engine is a r.v. The life time of Old component is with a fairly normal distribution $\mu_1 = 5000$ hours, and $\sigma_1 = 40$ hours; new component is with $\mu_2 = 5050$ hours, and $\sigma_2 = 30$ hours. We randomly select $n_1 = 16$ old components and $n_2 = 25$ new components from the process. What is the probabilities that the difference in the two sample means $\bar{X}_2 - \bar{X}_1$ is at least 25 hours?

Solution:

$$\mu_2 - \mu_1 = 5050 - 5000 = 50$$

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40^2}{16} + \frac{30^2}{25}} = \sqrt{136}$$

Since the distribution of life time of Old component is fairly normal, $n_1 = 16$ is ok to do CLT approximation; $n_2 = 25$ is close to 30, therefore, we can apply CLT to approximate the distribution of difference in sample means,

$$\bar{X}_2 - \bar{X}_1 \sim N(50, 136)$$

$$P(\bar{X}_2 - \bar{X}_1 \geq 25) = P(Z \geq \frac{25 - 50}{\sqrt{136}}) = P(Z \geq -2.14) = 0.9838$$

the probabilities that the difference in the two sample means $\bar{X}_2 - \bar{X}_1$ is at least 25 hours is 0.9838.

Bias of an estimator

- Unbiased estimator:
  ——-The point estimator of $\hat{\theta}$ is an unbiased estimator for the parameter θ if
  $$E(\hat{\theta}) = \theta$$
- Biased estimator:
  ——The point estimator of $\hat{\theta}$ is a biased estimator for the parameter θ if
  $$E(\hat{\theta}) \neq \theta$$
  —— $E(\hat{\theta}) - \theta$ is called the bias of the estimator of θ.
- Mean squared error:

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2] + (E(\hat{\theta}) - \theta)^2 \\ &= V(\hat{\theta}) + (\text{bias}[\hat{\theta}])^2 \end{aligned}$$

$$MSE(\hat{\theta}) = V(\hat{\theta}) + (\text{bias}[\hat{\theta}])^2.$$

———-If the estimator is unbiased, we usually select the estimator with the smallest variance.

———-If the estimator is biased, we usually select the estimator with the smallest **mean squared error**.