Point estimators

Chrysafis Vogiatzis

Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Lectures 15 and 16

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

- By looking at a sample, rather than the whole population, we save time and effort.
- By looking at a sample, rather than the whole population, we lose information.
- By looking at the numerical information of the sample, we are able to estimate the whole population.

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

1. By looking at a sample, rather than the whole population, we save time and effort.
2. By looking at a sample, rather than the whole population, we lose information.
3. By looking at the numerical information of the sample, we are able to recreate the whole population.

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

1. By looking at a sample, rather than the whole population, we save time and effort.

   True.

2. By looking at a sample, rather than the whole population, we lose information.

   Unfortunately, true.

3. By looking at the numerical information of the sample, we are able to recreate the whole population.

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

1. By looking at a sample, rather than the whole population, we save time and effort.

   True.

2. By looking at a sample, rather than the whole population, we lose information.

3. By looking at the numerical information of the sample, we are able to recreate the whole population.

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

1. By looking at a sample, rather than the whole population, we save time and effort.

   True.

2. By looking at a sample, rather than the whole population, we lose information.

   Unfortunately true.

3. By looking at the numerical information of the sample, we are able to recreate the whole population.

   We sure hope so.

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

1. By looking at a sample, rather than the whole population, we save time and effort.

   True.

2. By looking at a sample, rather than the whole population, we lose information.

   Unfortunately true.

3. By looking at the numerical information of the sample, we are able to recreate the whole population.

   We sure hope so.

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

1. By looking at a sample, rather than the whole population, we save time and effort.

   True.

2. By looking at a sample, rather than the whole population, we lose information.

   Unfortunately true.

3. By looking at the numerical information of the sample, we are able to recreate the whole population.

   We sure hope so.

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

1. By looking at a sample, rather than the whole population, we save time and effort.

   True.

2. By looking at a sample, rather than the whole population, we lose information.

   Unfortunately true.

3. By looking at the numerical information of the sample, we are able to recreate the whole population.

   We sure hope so.

Statistical inference

- What does a sample tell us for the whole population?
  - We interviewed 50 people about the next election. What do the results imply for the general election?
  - We picked a sample of 10 cars and performed a crash test. What do the observations imply for the whole production line?
  - We collected exit interview data from 100 alumni. What do their answers imply for the starting salary of our alumni?

Some observations:

1. By looking at a sample, rather than the whole population, we save time and effort.

   True.

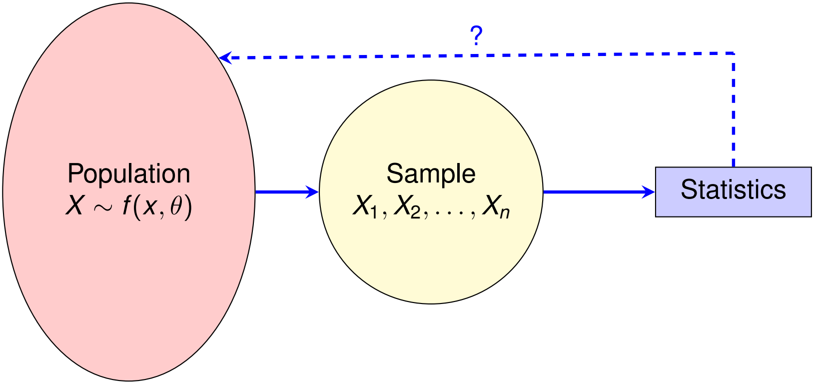
2. By looking at a sample, rather than the whole population, we lose information.

   Unfortunately true.

3. By looking at the numerical information of the sample, we are able to recreate the whole population.

   We sure hope so.

Theme



What is a statistic?

A statistic is *any* value obtained by random data.

- Statistics depend on the sample selected!
- Statistics are functions of the sample selected.
- Statistics are random variables.

What does the probability distribution of a sample depend on?

1. The distribution of the population.
2. The size of the sample selected.
3. The way the sample was selected.

This is referred to as a **sampling distribution**.

What is a statistic?

A statistic is *any* value obtained by random data.

- Statistics depend on the sample selected!
- Statistics are functions of the sample selected.
- Statistics are random variables.

What does the probability distribution of a sample depend on?

1. The distribution of the population.
2. The size of the sample selected.
3. The way the sample was selected.

This is referred to as a **sampling distribution**.

What is a statistic?

A statistic is *any* value obtained by random data.

- Statistics depend on the sample selected!
- Statistics are functions of the sample selected.
- Statistics are random variables.

What does the probability distribution of a sample depend on?

1. The distribution of the population.
2. The size of the sample selected.
3. The way the sample was selected.

This is referred to as a **sampling distribution**.

ILLINOIS

What is a statistic?

A statistic is *any* value obtained by random data.

- Statistics depend on the sample selected!
- Statistics are functions of the sample selected.
- Statistics are random variables.

What does the probability distribution of a sample depend on?

1. The distribution of the population.
2. The size of the sample selected.
3. The way the sample was selected.

This is referred to as a **sampling distribution**.

ILLINOIS

What is a statistic?

A statistic is *any* value obtained by random data.

- Statistics depend on the sample selected!
- Statistics are functions of the sample selected.
- Statistics are random variables.

What does the probability distribution of a sample depend on?

1. The distribution of the population.
2. The size of the sample selected.
3. The way the sample was selected.

This is referred to as a **sampling distribution**.

What is a statistic?

A statistic is *any* value obtained by random data.

- Statistics depend on the sample selected!
- Statistics are functions of the sample selected.
- Statistics are random variables.

What does the probability distribution of a sample depend on?

1. The distribution of the population.
2. The size of the sample selected.
3. The way the sample was selected.

This is referred to as a **sampling distribution**.

ILLINOIS

What is a statistic?

A statistic is *any* value obtained by random data.

- Statistics depend on the sample selected!
- Statistics are functions of the sample selected.
- Statistics are random variables.

What does the probability distribution of a sample depend on?

1. The distribution of the population.
2. The size of the sample selected.
3. The way the sample was selected.

This is referred to as a **sampling distribution**.

ILLINOIS

Point estimators

- Let X be a population following some pdf $f(x, \theta)$, with θ being some **unknown parameter** (could be a vector of multiple unknown parameters).
- Let $X_1, X_2, \dots, X_n$ be a random sample from X (identically distributed and independent random variables X_i) of size n.
- **Point estimator**: a statistic $\hat{\Theta}$ used to approximate θ.
  - Of course $\hat{\Theta}$ is a function of $X_1, X_2, \dots, X_n$: $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$.
  - $\hat{\Theta}$ is also a random variable.
- **Point estimate**: the actual numerical value $\hat{\theta}$ that the statistic has for a specific, given sample.
  - $\hat{\theta}$ is a number, specific to the sample observed.
- Say the full population is $X = \{10, 12, 13, 15, 16, 24\}$ (with a mean of 16, but assume we do not know that). The mean is θ.
- Now, we use a point estimator to estimate the mean. We pick a sample of 2 elements, and report their average. Then, $\hat{\Theta} = (X_1 + X_2)/2$.
- If we pick $X_1 = 10, X_2 = 16$, the average is 13, so we report $\hat{\theta} = 13$. On the other hand, had we picked 16, 24, we'd report $\hat{\theta} = 20$.

Point estimators

- Let X be a population following some pdf $f(x, \theta)$, with θ being some **unknown parameter** (could be a vector of multiple unknown parameters).
- Let $X_1, X_2, \ldots, X_n$ be a random sample from X (identically distributed and independent random variables X_i) of size n.
- **Point estimator**: a statistic $\hat{\Theta}$ used to approximate θ.
  - Of course $\hat{\Theta}$ is a function of $X_1, X_2, \ldots, X_n$: $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$.
  - $\hat{\Theta}$ is also a random variable.
- **Point estimate**: the actual numerical value $\hat{\theta}$ that the statistic has for a specific, given sample.
  - $\hat{\theta}$ is a number, specific to the sample observed.
- Say the full population is $X = \{10, 12, 13, 15, 16, 24\}$ (with a mean of 16, but assume we do not know that). The mean is θ.
- Now, we use a point estimator to estimate the mean. We pick a sample of 2 elements, and report their average. Then, $\hat{\Theta} = (X_1 + X_2)/2$.
- If we pick $X_1 = 10, X_2 = 16$, the average is 13, so we report $\hat{\theta} = 13$. On the other hand, had we picked 16, 24, we'd report $\hat{\theta} = 20$.

ILLINOIS

Point estimators

- Let X be a population following some pdf $f(x, \theta)$, with θ being some **unknown parameter** (could be a vector of multiple unknown parameters).
- Let $X_1, X_2, \ldots, X_n$ be a random sample from X (identically distributed and independent random variables X_i) of size n.
- **Point estimator**: a statistic $\hat{\Theta}$ used to approximate θ.
  - Of course $\hat{\Theta}$ is a function of $X_1, X_2, \ldots, X_n$: $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$.
  - $\hat{\Theta}$ is also a random variable.
- **Point estimate**: the actual numerical value $\hat{\theta}$ that the statistic has for a specific, given sample.
  - $\hat{\theta}$ is a number, specific to the sample observed.
- Say the full population is $X = \{10, 12, 13, 15, 16, 24\}$ (with a mean of 16, but assume we do not know that). The mean is θ.
- Now, we use a point estimator to estimate the mean. We pick a sample of 2 elements, and report their average. Then, $\hat{\Theta} = (X_1 + X_2)/2$.
- If we pick $X_1 = 10, X_2 = 16$, the average is 13, so we report $\hat{\theta} = 13$. On the other hand, had we picked 16, 24, we'd report $\hat{\theta} = 20$.

Point estimators

- Let X be a population following some pdf $f(x, \theta)$, with θ being some **unknown parameter** (could be a vector of multiple unknown parameters).
- Let $X_1, X_2, \ldots, X_n$ be a random sample from X (identically distributed and independent random variables X_i) of size n.
- **Point estimator**: a statistic $\hat{\Theta}$ used to approximate θ.
  - Of course $\hat{\Theta}$ is a function of $X_1, X_2, \ldots, X_n$: $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$.
  - $\hat{\Theta}$ is also a random variable.
- **Point estimate**: the actual numerical value $\hat{\theta}$ that the statistic has for a specific, given sample.
  - $\hat{\theta}$ is a number, specific to the sample observed.
- Say the full population is $X = \{10, 12, 13, 15, 16, 24\}$ (with a mean of 16, but assume we do not know that). The mean is θ.
- Now, we use a point estimator to estimate the mean. We pick a sample of 2 elements, and report their average. Then, $\hat{\Theta} = (X_1 + X_2)/2$.
- If we pick $X_1 = 10, X_2 = 16$, the average is 13, so we report $\hat{\theta} = 13$. On the other hand, had we picked 16, 24, we'd report $\hat{\theta} = 20$.

Point estimators

- Let X be a population following some pdf $f(x,\theta)$, with θ being some **unknown parameter** (could be a vector of multiple unknown parameters).
- Let $X_1, X_2, \ldots, X_n$ be a random sample from X (identically distributed and independent random variables X_i) of size n.
- **Point estimator**: a statistic $\hat{\Theta}$ used to approximate θ.
  - Of course $\hat{\Theta}$ is a function of $X_1, X_2, \ldots, X_n$: $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$.
  - $\hat{\Theta}$ is also a random variable.
- **Point estimate**: the actual numerical value $\hat{\theta}$ that the statistic has for a specific, given sample.
  - $\hat{\theta}$ is a number, specific to the sample observed.
- Say the full population is $X = \{10, 12, 13, 15, 16, 24\}$ (with a mean of 16, but assume we do not know that). The mean is θ.
- Now, we use a point estimator to estimate the mean. We pick a sample of 2 elements, and report their average. Then, $\hat{\Theta} = (X_1 + X_2)/2$.
- If we pick $X_1 = 10, X_2 = 16$, the average is 13, so we report $\hat{\theta} = 13$. On the other hand, had we picked 16, 24, we'd report $\hat{\theta} = 20$.

Point estimators

- Let X be a population following some pdf $f(x, \theta)$, with θ being some **unknown parameter** (could be a vector of multiple unknown parameters).
- Let $X_1, X_2, \ldots, X_n$ be a random sample from X (identically distributed and independent random variables X_i) of size n.
- **Point estimator**: a statistic $\hat{\Theta}$ used to approximate θ.
  - Of course $\hat{\Theta}$ is a function of $X_1, X_2, \ldots, X_n$: $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$.
  - $\hat{\Theta}$ is also a random variable.
- **Point estimate**: the actual numerical value $\hat{\theta}$ that the statistic has for a specific, given sample.
  - $\hat{\theta}$ is a number, specific to the sample observed.

- Say the full population is $X = \{10, 12, 13, 15, 16, 24\}$ (with a mean of 16, but assume we do not know that). The mean is θ.
- Now, we use a point estimator to estimate the mean. We pick a sample of 2 elements, and report their average. Then, $\hat{\Theta} = (X_1 + X_2)/2$.
- If we pick $X_1 = 10, X_2 = 16$, the average is 13, so we report $\hat{\theta} = 13$. On the other hand, had we picked 16, 24, we'd report $\hat{\theta} = 20$.

ILLINOIS

Point estimators

- Let X be a population following some pdf $f(x, \theta)$, with θ being some **unknown parameter** (could be a vector of multiple unknown parameters).
- Let $X_1, X_2, \ldots, X_n$ be a random sample from X (identically distributed and independent random variables X_i) of size n.
- **Point estimator**: a statistic $\hat{\Theta}$ used to approximate θ.
  - Of course $\hat{\Theta}$ is a function of $X_1, X_2, \ldots, X_n$: $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$.
  - $\hat{\Theta}$ is also a random variable.
- **Point estimate**: the actual numerical value $\hat{\theta}$ that the statistic has for a specific, given sample.
  - $\hat{\theta}$ is a number, specific to the sample observed.

- Say the full population is $X = \{10, 12, 13, 15, 16, 24\}$ (with a mean of 16, but assume we do not know that). The mean is θ.
- Now, we use a point estimator to estimate the mean. We pick a sample of 2 elements, and report their average. Then, $\hat{\Theta} = (X_1 + X_2)/2$.
- If we pick $X_1 = 10, X_2 = 16$, the average is 13, so we report $\hat{\theta} = 13$. On the other hand, had we picked 16, 24, we'd report $\hat{\theta} = 20$.

ILLINOIS

Point estimators

- Let X be a population following some pdf $f(x, \theta)$, with θ being some **unknown parameter** (could be a vector of multiple unknown parameters).
- Let $X_1, X_2, \ldots, X_n$ be a random sample from X (identically distributed and independent random variables X_i) of size n.
- **Point estimator**: a statistic $\hat{\Theta}$ used to approximate θ.
  - Of course $\hat{\Theta}$ is a function of $X_1, X_2, \ldots, X_n$: $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$.
  - $\hat{\Theta}$ is also a random variable.
- **Point estimate**: the actual numerical value $\hat{\theta}$ that the statistic has for a specific, given sample.
  - $\hat{\theta}$ is a number, specific to the sample observed.

- Say the full population is $X = \{10, 12, 13, 15, 16, 24\}$ (with a mean of 16, but assume we do not know that). The mean is θ.
- Now, we use a point estimator to estimate the mean. We pick a sample of 2 elements, and report their average. Then, $\hat{\Theta} = (X_1 + X_2)/2$.
- If we pick $X_1 = 10, X_2 = 16$, the average is 13, so we report $\hat{\theta} = 13$. On the other hand, had we picked 16, 24, we'd report $\hat{\theta} = 20$.

ILLINOIS

Commonly used point estimators

For a single population:

Parameters	Point estimators
Population mean μ	Sample average $\hat{\Theta} = \overline{x}$
Population variance σ^2	Sample variance $\hat{\Theta} = s^2$
Population proportion p	Sample proportion $\frac{\hat{n}}{n}$

For two populations:

Parameters	Point estimators
Difference in population means $\mu_1 - \mu_2$	Difference in sample averages $\hat{\Theta} = \overline{x}_1 - \overline{x}_2$
Ratio in population variances $\frac{\sigma_1^2}{\sigma_2^2}$	Ratio in sample variance $\frac{s_1^2}{s_2^2}$
Difference in population proportions $p_1 - p_2$	Difference in sample proportions $\hat{\Theta} = \frac{\hat{n}_1}{n_1} - \frac{\hat{n}_2}{n_2}$

Evaluating estimators

- **Bias** of point estimator $\hat{\Theta}$:

$$\mathit{bias}(\hat{\Theta}) = E\left[\hat{\Theta}\right] - \theta.$$

  - If $\mathit{bias}(\hat{\Theta}) = 0$, then $\hat{\Theta}$ is unbiased.
  - Bias is a measure of accuracy.

- **Standard error** of point estimator $\hat{\Theta}$:

$$SE\left[\hat{\Theta}\right] = \sqrt{Var\left[\hat{\Theta}\right]}.$$

  - We want this to be minimum.
  - SE is a measure of precision.

An estimator with zero bias and minimum variance among all other estimators is called an **minimum variance unbiased estimator**.

ILLINOIS

Evaluating estimators

- **Bias** of point estimator $\hat{\Theta}$:

$$bias(\hat{\Theta}) = E\left[\hat{\Theta}\right] - \theta.$$

  - If $bias(\hat{\Theta}) = 0$, then $\hat{\Theta}$ is unbiased.
  - Bias is a measure of accuracy.

- **Standard error** of point estimator $\hat{\Theta}$:

$$SE\left[\hat{\Theta}\right] = \sqrt{Var\left[\hat{\Theta}\right]}.$$

  - We want this to be minimum.
  - SE is a measure of precision.

An estimator with zero bias and minimum variance among all other estimators is called an **minimum variance unbiased estimator**.

ILLINOIS

Evaluating estimators

- **Bias** of point estimator $\hat{\Theta}$:

$$bias(\hat{\Theta}) = E\left[\hat{\Theta}\right] - \theta.$$

  - If $bias(\hat{\Theta}) = 0$, then $\hat{\Theta}$ is unbiased.
  - Bias is a measure of accuracy.

- **Standard error** of point estimator $\hat{\Theta}$:

$$SE\left[\hat{\Theta}\right] = \sqrt{Var\left[\hat{\Theta}\right]}.$$

  - We want this to be minimum.
  - SE is a measure of precision.

An estimator with zero bias and minimum variance among all other estimators is called an **minimum variance unbiased estimator**.

ILLINOIS

Evaluating estimators

- **Bias** of point estimator $\hat{\Theta}$:

$$bias(\hat{\Theta}) = E\left[\hat{\Theta}\right] - \theta.$$

  - If $bias(\hat{\Theta}) = 0$, then $\hat{\Theta}$ is unbiased.
  - Bias is a measure of accuracy.

- **Standard error** of point estimator $\hat{\Theta}$:

$$SE\left[\hat{\Theta}\right] = \sqrt{Var\left[\hat{\Theta}\right]}.$$

  - We want this to be minimum.
  - SE is a measure of precision.

An estimator with zero bias and minimum variance among all other estimators is called an **minimum variance unbiased estimator**.

ILLINOIS

Comparing estimators

So, we want minimum bias and variance. But how do we compare estimators based on these two numbers?

- **Mean square error** of point estimator $\hat{\Theta}$:

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

  - MSE quantifies both accuracy and precision.

- Given two estimators $\hat{\Theta}_1, \hat{\Theta}_2$ we compare them through their relative efficiency:

$$\text{Relative efficiency} = \frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_2\right)}.$$

  - Relative efficiency < 1 implies that $\hat{\Theta}_1$ is preferable.

ILLINOIS

Comparing estimators

So, we want minimum bias and variance. But how do we compare estimators based on these two numbers?

- **Mean square error** of point estimator $\hat{\Theta}$:

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

  - MSE quantifies both accuracy and precision.

- Given two estimators $\hat{\Theta}_1, \hat{\Theta}_2$ we compare them through their relative efficiency:

$$\text{Relative efficiency} = \frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_2\right)}.$$

  - Relative efficiency < 1 implies that $\hat{\Theta}_1$ is preferable.

Comparing estimators

So, we want minimum bias and variance. But how do we compare estimators based on these two numbers?

- **Mean square error** of point estimator $\hat{\Theta}$:

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

  - MSE quantifies both accuracy and precision.

- Given two estimators $\hat{\Theta}_1, \hat{\Theta}_2$ we compare them through their relative efficiency:

$$\text{Relative efficiency} = \frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_2\right)}.$$

  - Relative efficiency < 1 implies that $\hat{\Theta}_1$ is preferable.

ILLINOIS

Example

Assume a population with mean μ and variance σ^2. As the mean is unknown you decide to use the following three approaches to estimate it:

1. *Get the average of 3 observations.*
2. *Get 3 observations and calculate $\frac{2 \cdot X_1 + X_2 - X_3}{2}$.*
3. *Get 3 observations and calculate $2X_1 + X_2 - X_3$.*

Which one is the best among them?

1. $\hat{\Theta}_1 = \frac{X_1 + X_2 + X_3}{3}$:

$$E\left[\hat{\Theta}_1\right] = E\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3}\left(E\left[X_1\right] + E\left[X_2\right] + E\left[X_3\right]\right) = $$

$$= \frac{1}{3}(\mu + \mu + \mu) = \mu \implies bias(\hat{\Theta}_1) = 0$$

$$Var\left[\hat{\Theta}_1\right] = Var\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{9}\left(Var\left[X_1\right] + Var\left[X_2\right] + Var\left[X_3\right]\right) = $$

$$= \frac{1}{9} 3\sigma^2 = \frac{\sigma^2}{3}$$

Combining, $MSE(\hat{\Theta}_1) = \frac{\sigma^2}{3} + 0 = \frac{\sigma^2}{3}$.

ILLINOIS

Example

Assume a population with mean μ and variance σ^2. As the mean is unknown you decide to use the following three approaches to estimate it:

1. *Get the average of 3 observations.*
2. *Get 3 observations and calculate* $\frac{2 \cdot X_1 + X_2 - X_3}{2}$.
3. *Get 3 observations and calculate* $2X_1 + X_2 - X_3$.

Which one is the best among them?

1. $\hat{\Theta}_1 = \frac{X_1+X_2+X_3}{3}$:

$$E\left[\hat{\Theta}_1\right] = E\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3}\left(E\left[X_1\right] + E\left[X_2\right] + E\left[X_3\right]\right) =$$

$$= \frac{1}{3}\left(\mu + \mu + \mu\right) = \mu \implies bias(\hat{\Theta}_1) = 0$$

$$Var\left[\hat{\Theta}_1\right] = Var\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{9}\left(Var\left[X_1\right] + Var\left[X_2\right] + Var\left[X_3\right]\right) =$$

$$= \frac{1}{9} 3\sigma^2 = \frac{\sigma^2}{3}$$

Combining, $MSE(\hat{\Theta}_1) = \frac{\sigma^2}{3} + 0 = \frac{\sigma^2}{3}$.

ILLINOIS

Example

Assume a population with mean μ and variance σ^2. As the mean is unknown you decide to use the following three approaches to estimate it:

1. *Get the average of 3 observations.*
2. *Get 3 observations and calculate* $\frac{2 \cdot X_1 + X_2 - X_3}{2}$.
3. *Get 3 observations and calculate* $2X_1 + X_2 - X_3$.

Which one is the best among them?

1. $\hat{\Theta}_1 = \frac{X_1 + X_2 + X_3}{3}$:

$$E\left[\hat{\Theta}_1\right] = E\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3}\left(E\left[X_1\right] + E\left[X_2\right] + E\left[X_3\right]\right) =$$

$$= \frac{1}{3}\left(\mu + \mu + \mu\right) = \mu \implies bias(\hat{\Theta}_1) = 0$$

$$Var\left[\hat{\Theta}_1\right] = Var\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{9}\left(Var\left[X_1\right] + Var\left[X_2\right] + Var\left[X_3\right]\right) =$$

$$= \frac{1}{9} 3\sigma^2 = \frac{\sigma^2}{3}$$

Combining, $MSE(\hat{\Theta}_1) = \frac{\sigma^2}{3} + 0 = \frac{\sigma^2}{3}$.

ILLINOIS

Similarly, we calculate

2. $\hat{\Theta}_2 = \frac{2 \cdot X_1 + X_2 - X_3}{2}$:

$$E\left[\hat{\Theta}_2\right] = E\left[\frac{2 \cdot X_1 + X_2 - X_3}{2}\right] = \frac{2\mu + \mu - \mu}{2} = \mu \implies$$

$$\implies bias\left(\hat{\Theta}_2\right) = 0$$

$$Var\left[\hat{\Theta}_2\right] = Var\left[\frac{2 \cdot X_1 + X_2 - X_3}{2}\right] = Var\left[X_1\right] + \frac{1}{4}Var\left[X_2\right] + \frac{1}{4}Var\left[X_3\right] =$$

$$= \sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \implies Var\left[\hat{\Theta}_2\right] = \frac{3}{2}\sigma^2.$$

With $MSE(\hat{\Theta}_2) = \frac{3\sigma^2}{2} + 0 = \frac{3\sigma^2}{2}$.

3. $\hat{\Theta}_3 = 2 \cdot X_1 + X_2 - X_3$:

$$E\left[\hat{\Theta}\right] = E\left[2 \cdot X_1 + X_2 - X_3\right] = 2\mu + \mu - \mu = 2\mu \implies bias\left(\hat{\Theta}_3\right) = \mu$$

$$Var\left[\hat{\Theta}_3\right] = Var\left[2 \cdot X_1 + X_2 - X_3\right] = 4\sigma^2 + \sigma^2 + \sigma^2 \implies$$

$$\implies Var\left[\hat{\Theta}_3\right] = 6\sigma^2.$$

And $MSE(\hat{\Theta}_3) = 6\sigma^2 + \mu^2$.

Similarly, we calculate

2. $\hat{\Theta}_2 = \frac{2 \cdot X_1 + X_2 - X_3}{2}$:

$$E\left[\hat{\Theta}_2\right] = E\left[\frac{2 \cdot X_1 + X_2 - X_3}{2}\right] = \frac{2\mu + \mu - \mu}{2} = \mu \implies$$

$$\implies bias\left(\hat{\Theta}_2\right) = 0$$

$$Var\left[\hat{\Theta}_2\right] = Var\left[\frac{2 \cdot X_1 + X_2 - X_3}{2}\right] = Var\left[X_1\right] + \frac{1}{4} Var\left[X_2\right] + \frac{1}{4} Var\left[X_3\right] =$$

$$= \sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \implies Var\left[\hat{\Theta}_2\right] = \frac{3}{2}\sigma^2.$$

With $MSE(\hat{\Theta}_2) = \frac{3\sigma^2}{2} + 0 = \frac{3\sigma^2}{2}$.

3. $\hat{\Theta}_3 = 2 \cdot X_1 + X_2 - X_3$:

$$E\left[\hat{\Theta}\right] = E\left[2 \cdot X_1 + X_2 - X_3\right] = 2\mu + \mu - \mu = 2\mu \implies bias\left(\hat{\Theta}_3\right) = \mu$$

$$Var\left[\hat{\Theta}_3\right] = Var\left[2 \cdot X_1 + X_2 - X_3\right] = 4\sigma^2 + \sigma^2 + \sigma^2 \implies$$

$$\implies Var\left[\hat{\Theta}_3\right] = 6\sigma^2.$$

And $MSE(\hat{\Theta}_3) = 6\sigma^2 + \mu^2$.

Similarly, we calculate

2. $\hat{\Theta}_2 = \frac{2 \cdot X_1 + X_2 - X_3}{2}$:

$$E\left[\hat{\Theta}_2\right] = E\left[\frac{2 \cdot X_1 + X_2 - X_3}{2}\right] = \frac{2\mu + \mu - \mu}{2} = \mu \implies$$

$$\implies bias\left(\hat{\Theta}_2\right) = 0$$

$$Var\left[\hat{\Theta}_2\right] = Var\left[\frac{2 \cdot X_1 + X_2 - X_3}{2}\right] = Var\left[X_1\right] + \frac{1}{4}Var\left[X_2\right] + \frac{1}{4}Var\left[X_3\right] =$$

$$= \sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \implies Var\left[\hat{\Theta}_2\right] = \frac{3}{2}\sigma^2.$$

With $MSE(\hat{\Theta}_2) = \frac{3\sigma^2}{2} + 0 = \frac{3\sigma^2}{2}$.

3. $\hat{\Theta}_3 = 2 \cdot X_1 + X_2 - X_3$:

$$E\left[\hat{\Theta}\right] = E\left[2 \cdot X_1 + X_2 - X_3\right] = 2\mu + \mu - \mu = 2\mu \implies bias\left(\hat{\Theta}_3\right) = \mu$$

$$Var\left[\hat{\Theta}_3\right] = Var\left[2 \cdot X_1 + X_2 - X_3\right] = 4\sigma^2 + \sigma^2 + \sigma^2 \implies$$

$$\implies Var\left[\hat{\Theta}_3\right] = 6\sigma^2.$$

And $MSE(\hat{\Theta}_3) = 6\sigma^2 + \mu^2$.

A quick review

Let us review very quickly the notions we have seen in this lecture:

- **Random sample**: $X_1, X_2, \ldots, X_n$ each independent and from the same population with mean μ and variance σ^2.
  - $E[X_i] = E[X] = \mu$.
  - $Var[X_i] = Var[X] = \sigma^2$.
- **Statistic**: any function of a random variable.
- **Sampling distribution**: the distribution of a statistic.
- **Point estimator** $\hat{\Theta}$: a statistic to estimate or approximate an unknown parameter θ.
- **Bias**: $E\left[\hat{\Theta}\right] - \theta$. we want this to be zero.
- **Standard error**: $\sqrt{Var\left[\hat{\Theta}\right]}$. we want this to be small.
- **Minimum variance unbiased estimator**: an estimator $\hat{\Theta}$ with zero bias and minimum variance.

ILLINOIS

How to get a good estimator?

Clearly, for some parameters we have excellent intuition:

- sample average in lieu of true population mean.
- sample variance in lieu of population variance.

But: in some cases, we have no easy idea of what would work.

Consider the following examples:

- How to estimate the rate of an exponential distribution?

  "We know the time between accidents in a factory is exponentially distributed. How do we find out what the rate is?"

- How to estimate the probability of a binomial distribution?

  "We know the number of students graduating from the Grainger College of Engineering in 4 years is binomially distributed. How do we find out what the probability of graduation in 4 years is?"

ILLINOIS

How to get a good estimator?

Clearly, for some parameters we have excellent intuition:

- sample average in lieu of true population mean.
- sample variance in lieu of population variance.

But: in some cases, we have no easy idea of what would work.

Consider the following examples:

- How to estimate the rate of an exponential distribution?

  "We know the time between accidents in a factory is exponentially distributed. How do we find out what the rate is?"

- How to estimate the probability of a binomial distribution?

  "We know the number of students graduating from the Grainger College of Engineering in 4 years is binomially distributed. How do we find out what the probability of graduation in 4 years is?"

ILLINOIS

How to get a good estimator?

Clearly, for some parameters we have excellent intuition:

- sample average in lieu of true population mean.
- sample variance in lieu of population variance.

But: in some cases, we have no easy idea of what would work.

Consider the following examples:

- How to estimate the rate of an exponential distribution?

  "We know the time between accidents in a factory is exponentially distributed. How do we find out what the rate is?"

- How to estimate the probability of a binomial distribution?

  "We know the number of students graduating from the Grainger College of Engineering in 4 years is binomially distributed. How do we find out what the probability of graduation in 4 years is?"

ILLINOIS

How to get a good estimator?

Clearly, for some parameters we have excellent intuition:

- sample average in lieu of true population mean.
- sample variance in lieu of population variance.

But: in some cases, we have no easy idea of what would work.

Consider the following examples:

- How to estimate the rate of an exponential distribution?

  "We know the time between accidents in a factory is exponentially distributed. How do we find out what the rate is?"

- How to estimate the probability of a binomial distribution?

  "We know the number of students graduating from the Grainger College of Engineering in 4 years is binomially distributed. How do we find out what the probability of graduation in 4 years is?"

ILLINOIS

How to get a good estimator?

The most commonly used methods of point estimation are:

1. The method of moments (or, moment matching);

   Lecture 17.

2. Maximum likelihood estimation (MLE);

   Lecture 18.

3. Bayesian estimation.

   Lecture 19.

How to get a good estimator?

The most commonly used methods of point estimation are:

1. The method of moments (or, moment matching);
   Lecture 17.
2. Maximum likelihood estimation (MLE);
   Lecture 18.
3. Bayesian estimation.
   Lecture 19.

ILLINOIS