Chapter 5: Special topics

Yonghyun Kwon
Department of Mathematics, Korea Military Academy

Sampling distribution

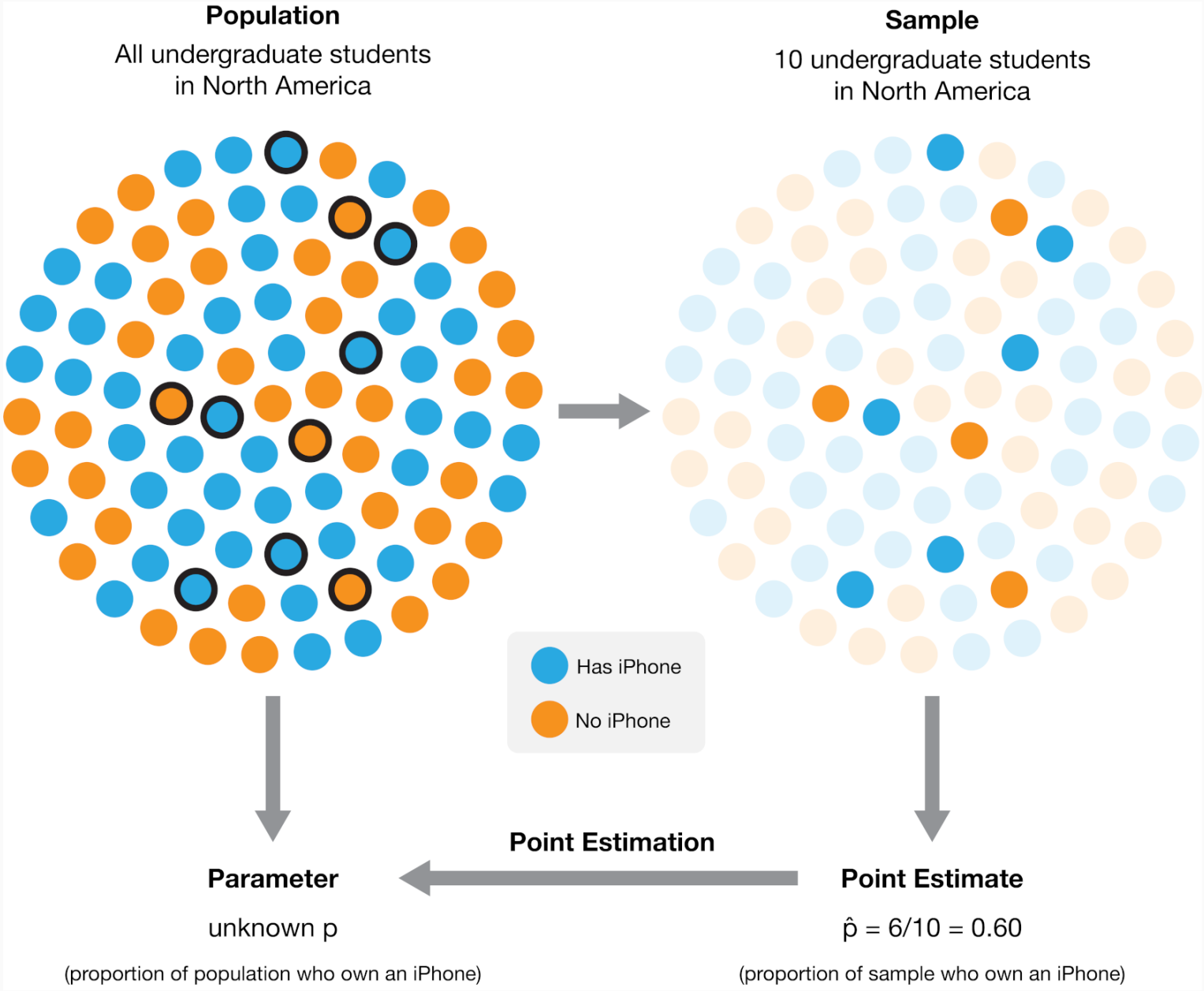
Population and Sample

Example: Estimating the average lifetime of company light bulbs.

- *Population*: The entire set of objects under study.
- *Parameter*: A numerical value describing a characteristic of the population (e.g., population mean, population variance, population proportion).
- *Sample*: A subset drawn from the population.

Key Idea

Because testing the entire population may be impossible (cost, time), we use samples to estimate population characteristics.



https://datasciencebook.ca/img/inference/population_vs_sample.png

Random Sample

Suppose a population follows some probability distribution F.

- A *random sample* of size n is a collection of n independent random variables

$$X_1, X_2, \ldots, X_n \sim F$$

drawn from the same distribution.

Example: Selecting 10 bulbs independently and recording their lifetimes. Then $X_1, \ldots, X_{10}$ are *independent and identically distributed(i.i.d.)* random variables with the same distribution as the population lifetime.

Statistic and Sampling Distribution

- A *statistic* is a function of the sample that does not involve unknown parameters.
- The probability distribution of a statistic is called a *sampling distribution*.
- Since a statistic is computed from a random sample, it is itself a random variable.

Examples of statistic:

$$\text{(Sample mean)} \quad : \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

Practice:

Let X be a discrete random variable taking values $0, 1, 2$ with probabilities 0.3, 0.6, and 0.1. Taking a random sample of size $n = 2$, find the distribution of $\bar{X} = (X_1 + X_2)/2$.

X	0	1	2
$f(X)$	0.3	0.6	0.1

Population distribution of X

X_1 \ X_2	0	1	2
0	0.09	0.18	0.03
1	0.18	0.36	0.06
2	0.03	0.06	0.01

Joint distribution of (X_1, X_2)

$\bar{X}$	0	0.5	1	1.5	2
$P(\bar{X})$	0.09	0.36	0.42	0.12	0.01

Sampling distribution of $\bar{X}$

Expectation and Variance of the Sample Mean

Expectation and Variance of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. Then

$$E[\bar{X}] = \mu, \qquad Var(\bar{X}) = \frac{\sigma^2}{n},$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Sample Mean from a Normal Population

Recall: a linear combination of two RVs

If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent,
$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$

If $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$,

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

$E(\bar{X}) = \mu$

and standarizing $\bar{X}$ gives

$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

$$Z = \frac{\bar{X} - \text{mean}}{SE} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Practice

The fuel efficiency (km/L) of a new car model follows $X \sim N(18, 2^2)$. Suppose we randomly select $n = 4$ cars and record their fuel efficiency $X_1, \ldots, X_4$. Compute $P(\bar{X} \leq 17)$.

- Since $X_i \sim N(18, 2^2)$, $E(\bar{X}) = 18$, $Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{4}{4} = 1$

$$\bar{X} \sim N(18, 1^2)$$

- Standardizing:

$$P(\bar{X} \leq 17) = P\left(\frac{\bar{X} - 18}{1} \leq \frac{17 - 18}{1}\right)$$

$$= P(Z \leq -1) = 0.1587$$

Central Limit Theorem (CLT)

Even if the population distribution is *not normal*: $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$

Central Limit Theorem (CLT)

For a random sample $X_1, \ldots, X_n$ from a population with mean μ and variance $\sigma^2 < \infty$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$$

- When n is large, $\bar{X}$ is approximately normal, regardless of the population distribution.

Practice

The monthly salary (in 10,000 KRW) of new 4-year college graduates has mean 266 and standard deviation 15. Suppose we take a random sample of $n = 81$ graduates. What is the probability that their average salary exceeds 270?

$E(\bar{X}) = 266$, $Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{15^2}{81}$

Since $n = 81$ is large enough, we can apply the CLT.

$$\bar{X} \sim N\left(266, \left(\frac{15}{9}\right)^2\right)$$

$$P(\bar{X} \geq 270) = P\left(Z \geq \frac{270 - 266}{5/3}\right) = P(Z \geq 2.4) \approx 0.0082.$$

Sample Variance Distribution

- To make inference about population variance σ^2, we use the sample variance

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

- Its distribution is related to the *chi-squared distribution*.

Chi-Squared Distribution

Chi-squared Distribution

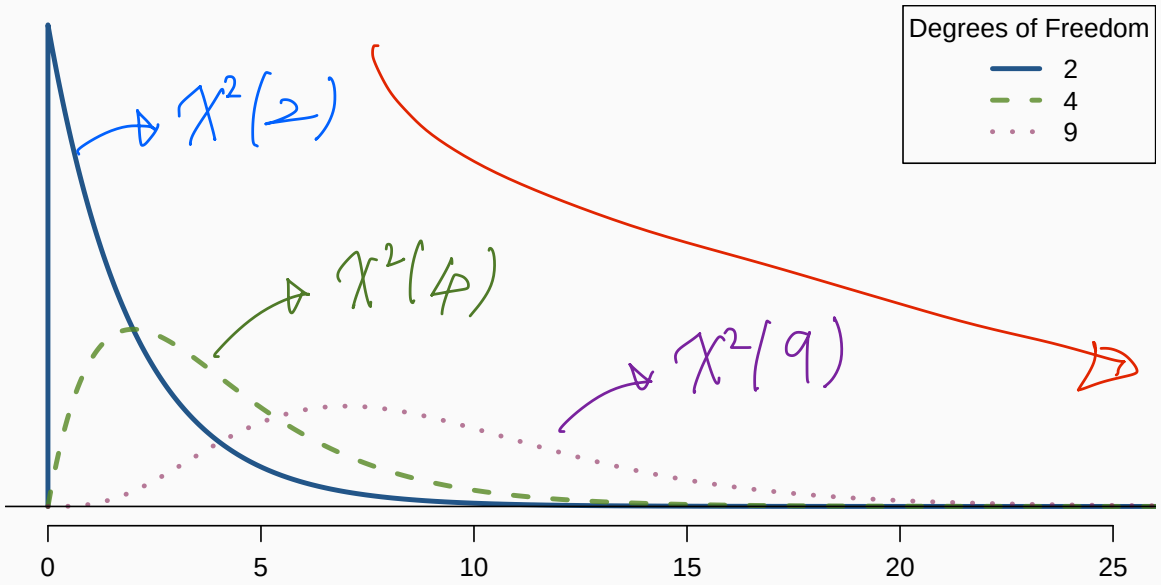
If $Z_1, Z_2, \cdots, Z_n \sim N(0,1)$ independently, $V := Z_1^2 + Z_2^2, \cdots + Z_n^2$ follows chi-square distribution with *degrees of freedom* n.

$$V \sim \chi^2(n),$$

- For small n, the distribution is *skewed to the right.*
- As n increases, it becomes more symmetric and approaches normal.

Chi-Square Distribution

Which of the following is false about chi-square distribution?



As the df increases,

(a) the center of the χ^2 distribution increases as well. ○

(b) the variability of the χ^2 distribution increases as well. ○

(c) the shape of the χ^2 distribution becomes more skewed. ✗

Properties of the Chi-Squared Distribution

Additivity of the Chi-Squared Distribution

If $V_1 \sim \chi^2(n_1)$ and $V_2 \sim \chi^2(n_2)$ independently, then

$$V_1 + V_2 \sim \chi^2(n_1 + n_2).$$

Relationship to Normal distribution

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

$$Z^2 \sim \chi^2(1).$$

Distribution of the Sample Variance

Distribution of the Sample Variance

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently. Then:

- $\bar{X}$ and S^2 are independent.
- The sample variance satisfies

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Note: This is fundamental for building confidence intervals and hypothesis tests about σ^2.

Distribution of the Sample Proportion

For a Bernoulli population with success probability p:

$$X_i \sim B(1,p), \quad i = 1,\ldots,n, \text{ independently,}$$

the sample proportion is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Since $\sum_{i=1}^{n} X_i \sim B(n,p)$, $E(\hat{p}) = p$ and $Var(\hat{p}) = \dfrac{p(1-p)}{n}$.

By the CLT, when n is large,

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right), \text{ or } \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \dot{\sim} N(0,1).$$

Practice

In a city, Party A's support rate is $p = 0.6$. A random sample of $n = 30$ voters is surveyed. What is the probability that the sample proportion $\hat{p}$ is less than 0.5?

The expectation and variance of the sample proportion $\hat{p}$ are

$$E(\hat{p}) = 0.6 \quad Var(\hat{p}) = \frac{p(1-p)}{n} = \frac{0.6 \times 0.4}{30} = 0.008$$

Since $np = 18 \geq 10$ and $n(1-p) = 12 \geq 10$, we can apply CLT:

$$\hat{p} \dot{\sim} N(0.6, \ 0.008)$$

Therefore,

$$P(\hat{p} \leq 0.5) = P\left(\frac{\hat{p} - 0.6}{\sqrt{0.008}} \leq \frac{0.5 - 0.6}{\sqrt{0.008}}\right) = P(Z \leq -1.12) \approx 0.1314.$$

Point estimation

Point Estimation

- *Point estimation* is the process of using sample information to estimate of an unknown population parameter.
- Statistics such as the sample mean, sample variance, and sample proportion are used as point estimators.
- *Estimator and Estimate:*
  - A statistic $\hat{\theta}(X_1, \ldots, X_n)$ used to estimate a parameter θ is called an *estimator*.
  - The computed value $\hat{\theta}(x_1, \ldots, x_n)$ from observed data $(x_1, \ldots, x_n)$ is called the *estimate*.
- In general, an estimator of the parameter θ is denoted by $\hat{\theta}$.

Example: Estimator and Estimate

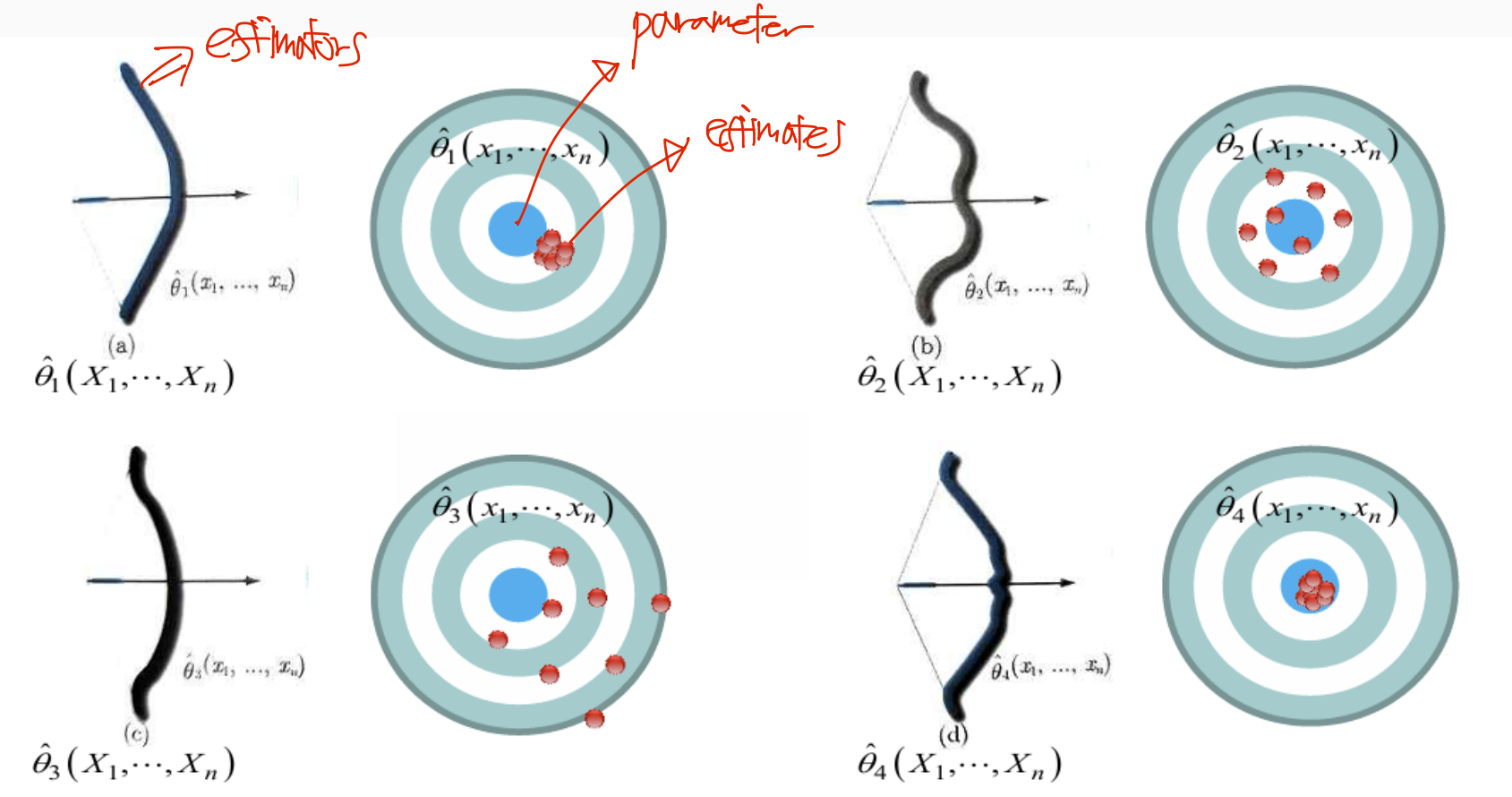
Suppose $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2.

	Mean	Variance
Parameter	μ	σ^2
Estimator	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Estimate	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Plugging in observed data $x_1, \ldots, x_n$ gives the estimate.

Properties of Estimators

Multiple estimators can be used to estimate the same parameter.



Desirable properties: *Unbiasedness*, *Efficiency*, *Consistency*

Unbiasedness

An estimator $\hat{\theta}$ of parameter θ is *unbiased* if

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

Its *bias* is defined as $B(\hat{\theta}) = E(\hat{\theta}) - \theta.$

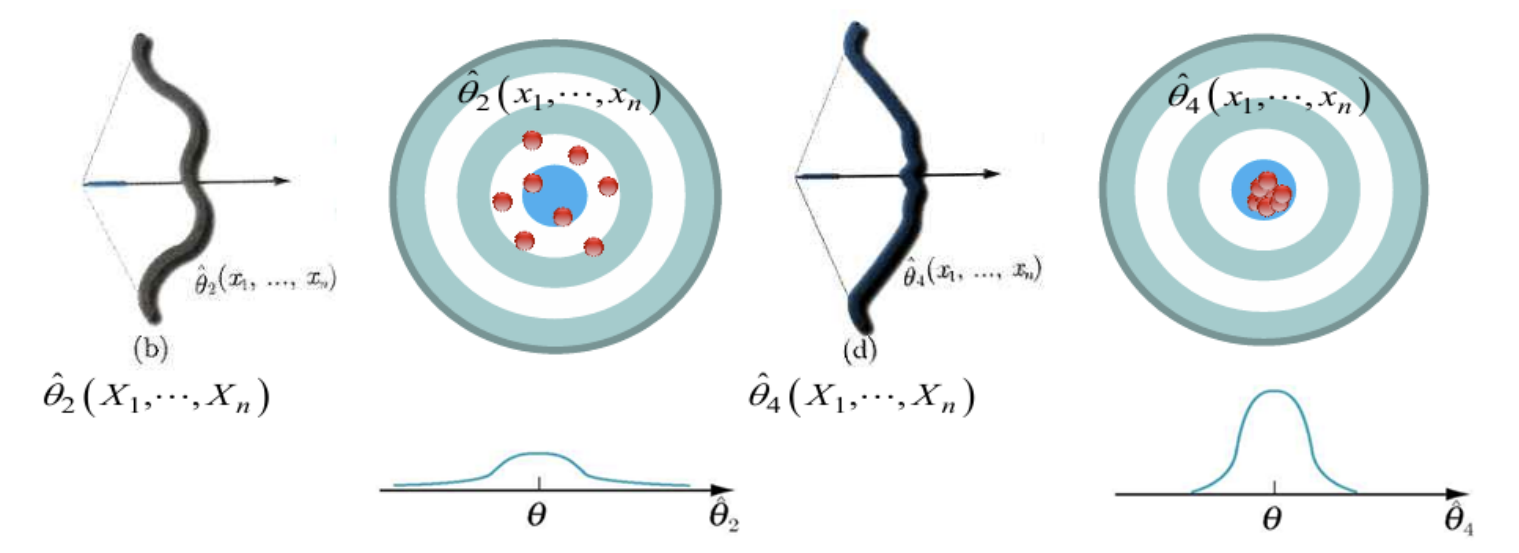
Suppose $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2. Show that the sample mean $\bar{X}$ is an unbiased estimator of μ.

From slide 6,

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Efficiency

Among unbiased estimators, the one with smaller variance is preferred.



Given two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of the same parameter θ, if

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

then $\hat{\theta}_1$ is said to be more *efficient*.

Example: Efficiency

Suppose X_1, X_2, X_3 a random sample from a population with mean μ and variance σ^2. Consider two estimators of μ:

$$T_1 = \frac{X_1 + 2X_2 + X_3}{4}, \qquad T_2 = \frac{X_1 + X_2 + X_3}{3}.$$

Which estimator is more efficient?

(independence)

$$Var(T_1) = \frac{1}{16}Var(X_1) + \frac{4}{16}Var(X_2) + \frac{1}{16}Var(X_3) = \frac{3}{8}\sigma^2$$

($Var(X_1) = \sigma^2$, $Var(X_2) = \sigma^2$, $Var(X_3) = \sigma^2$)

$$Var(T_2) = \frac{1}{9}(Var(X_1) + Var(X_2) + Var(X_3)) = \frac{1}{3}\sigma^2.$$

Since $\frac{3}{8}\sigma^2 > \frac{1}{3}\sigma^2$, T_2 is more efficient than T_1

Consistency

Consistency

An estimator $\hat{\theta}$ of parameter θ is *consistent* if

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta \text{ and } \lim_{n\to\infty} Var(\hat{\theta}) = 0$$

As the sample size increases, a consistent estimator *converges* in probability to the true parameter.

Example: Sample Mean

Properties of the Sample Mean

If $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2, then the sample mean $\bar{X}$ is both an *unbiased* and *consistent* estimator of μ.

Unbiasedness follows directly from slide 21. Also,

$$E(\bar{X}) = \mu, \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

gives

$$\lim_{n\to\infty} E(\bar{X}) = \lim_{n\to\infty} \mu = \mu, \qquad \lim_{n\to\infty} Var(\bar{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

Hence $\bar{X}$ is consistent for μ.

Example: Sample Variance

Properties of the Sample Variance

If $X_1, \ldots, X_n$ is a random sample from a population with variance σ^2, then the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is both an *unbiased* and *consistent* estimator of σ^2.

To check S^2 is unbiased for σ^2, we show that $E[S^2] = \sigma^2$:

$$
\begin{aligned}
(n-1)E[S^2] &= E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \left(\sum_{i=1}^{n} E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2]\right) \qquad (1) \\
&= \left(n\sigma^2 - n \cdot \frac{\sigma^2}{n}\right) = (n-1)\sigma^2,
\end{aligned}
$$

and (1) holds because

$$
\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\underbrace{\sum_{i=1}^{n}(X_i - \bar{X})}_{0}.
$$

Example: Sample Proportion

Properties of the Sample Proportion

If $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ independently, the sample proportion

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is both an *unbiased* and *consistent* estimator of p.

Recall that

$$E(\hat{p}) = p \qquad Var(\hat{p}) = \frac{p(1-p)}{n}$$

from slide 16. Thus, $\hat{p}$ is consistent for p. Also,

$$\lim_{n\to\infty} E(\hat{p}) = \lim_{n\to\infty} p = p, \qquad \lim_{n\to\infty} Var(\hat{p}) = \lim_{n\to\infty} \frac{p(1-p)}{n} = 0$$

Hence $\hat{p}$ is consistent for p.

Exercises

1. An employee at a call center takes between 0.5 and 6.5 minutes to respond to a phone call, and this response time is assumed to be uniformly distributed. If the response times of 60 independent calls handled by this employee are observed, approximate the probability that the average response time is at least 4 minutes.

Exercises

2. When Team A and Team B play a game in a sport, probabilities that Team A wins or loses are 0.6 and 0.4, respectively. Each outcome gives 3 points for a win and 0 points for a loss. Let $X_1, X_2, \ldots, X_n$ denote the points earned by Team A in n independent games (assume independence across games).

(1) Find the joint probability mass function of X_1 and X_2.

(2) Let $\bar{X}_2$ be the average score over 2 games. Find the probability mass function of $\bar{X}_2$.

(3) Find the expectation $E[\bar{X}_2]$ and variance $Var(\bar{X}_2)$.

(4) Suppose the two teams play 54 games. Approximate the probability that Team A's average score is at least 2.

Exercises

3. Suppose we obtain a random sample $X_1, X_2, \ldots, X_n$ from a population with probability density function

$$f(x) = \begin{cases} \frac{1}{3\theta}, & -\theta < x < 2\theta, \ (\theta > 0), \\ 0, & \text{otherwise.} \end{cases}$$

(1) Find the constant k such that the estimator $k\bar{X}$ is an unbiased estimator for θ, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(2) For the value of k obtained in (1), show that $k\bar{X}$ is a consistent estimator for θ.

Exercises

4. Suppose $X_1, X_2, \ldots, X_n$ are obtained from a Bernoulli distribution $B(1, p)$.

(1) Consider the estimators

$$\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{p}_2 = \frac{1}{2}(X_1 + X_2).$$

Determine whether each estimator is unbiased for p.

(2) Between $\hat{p}_1$ and $\hat{p}_2$, determine which estimator is more efficient (assume $n > 2$).

(3) Show that the estimator $\hat{p}_1$ is a consistent estimator for p.

Exercises

5. Suppose we obtain a random sample $X_1, X_2, \ldots, X_n$ from the population with density

$$f(x) = \begin{cases} \dfrac{1}{\theta+1} e^{-x/(\theta+1)}, & x > 0,\ \theta > -1, \\ 0, & \text{otherwise.} \end{cases}$$

(1) Find the expectation $E(X_1)$ and variance $Var(X_1)$.

(2) Consider the estimators

$$T_1 = X_1 - 1, \qquad T_2 = \bar{X} - 1 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - 1, \quad (n \geq 2).$$

Determine whether each is unbiased for θ.

(3) Between T_1 and T_2, which estimator is more efficient?

(4) Determine whether the more efficient estimator in (3) is consistent for θ.