ISE 315: Engineering Statistics

Lecture 2: Point Estimators and Sampling Distributions

Instructor: Mansur M. Arief, PhD
Industrial and Systems Engineering, KFUPM

Office: 22-219 — Email: mansur.arief@kfupm.edu.sa

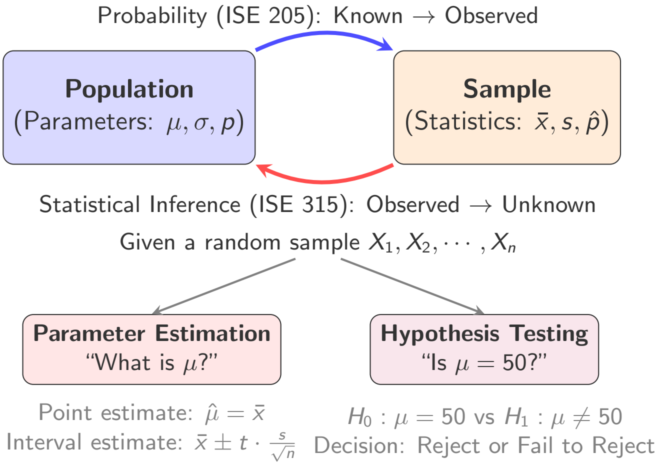
Lecture 2

Estimation: Point Estimators and Sampling Distributions

Lecture 2 Outline (Handbook Chapter 2)

- Point estimation (2.1)
- Sampling distributions and the Central Limit Theorem (2.8)
- Quality of estimators: bias, variance, standard error, MSE (Procedure 2.1)
- (Next lecture) Practice problems on estimators

Chapter 2 Concepts



Key Terminologies

1. **Random Sample (of size n):**
   - $X_1, X_2, \ldots, X_n$ are **independent**
   - Each X_i has a **same distribution**
   - Observed values: $x_1, x_2, \ldots, x_n$
2. **Statistic:**
   - Any function of observed samples
   - E.g: $\bar{X}$ ("X bar"), S^2, mode, etc.
   - A statistic is a **random variable**
3. **Point Estimator:**
   - A statistic $\hat{\theta}$ ("theta hat") used to estimate an unknown parameter θ ("theta"):
     $$\hat{\theta} = f(x_1, x_2, \ldots, x_n) \tag{2.1}$$
     - $\hat{\mu} = \bar{X}$ (sample average) estimates μ ("mu")
     - $\hat{\sigma}^2 = S^2$ (sample variance) estimates σ^2 ("sigma squared")

Key Terminologies

4. Sampling Distribution:

- Probability distribution of a statistic
  - Not about the population distribution of X_i's
  - But the sampling distribution (of the statistic), e.g $\bar{X}$
- Why do statistics vary?
  - Because different set of samples yield different statistics
- What affects the properties of the sampling distribution?
  - Depends on the statistic formula, sample size n, etc.

Different samples yield different estimates—this variability is what we study!

Key Terminologies

5. Properties of Estimators

- **Biased/unbiased:** Is the mean **equal** to the true parameter?
- **Low/high bias:** How much is it **off** from the true parameter? (2.3):

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

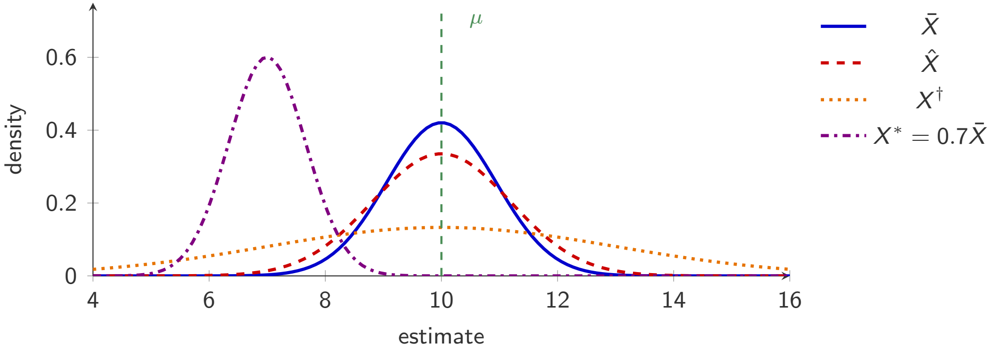
- **Low/high variance:** How much does it **vary** from sample to sample?
- **Rate of shrinkage:** How **quickly** does it converge as n increases?
- **Low/high Mean Squared Error (MSE):** Combines variance & bias (2.6):

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \text{Bias}^2$$

Ideal estimator: Low MSE (low variance and low bias)

Some estimators are better than others

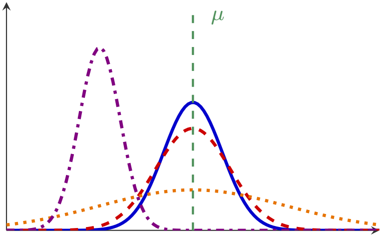
- Consider four estimators for normal population mean $\mu = 10$, $\sigma = 3$, $n = 10$:



Which one would you choose and why?

Some statistics are better than others

- What you have seen is the sampling distribution of
  - Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - Half-range: $\hat{X} = \frac{\max(X_i)+\min(X_i)}{2}$
  - First sample: $X^{\dagger} = X_1$
  - Shrinkage: $X^{*} = 0.7\bar{X}$
- Which one(s) are unbiased?
- Which one(s) has smallest variance?
- Which one has lowest MSE for $n = 1$? For larger n?

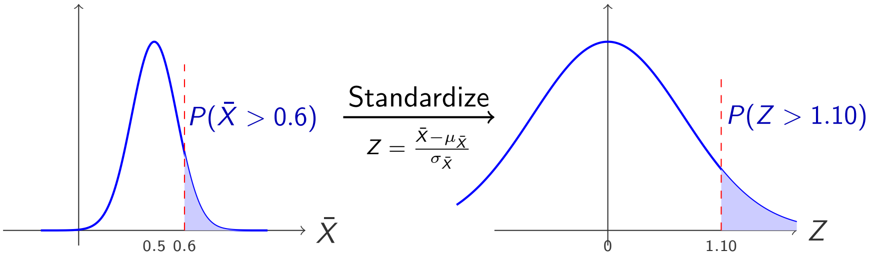


Notes about Estimator Properties

- About **sampling distribution**, NOT **population distribution**!
- The sampling dist **varies** with samples, but the population dist is **fixed**
- Example: an LLM eval assigns each response a score in $[0, 1]$; scores behave like $X_i \sim \text{Uniform}(0, 1)$
  - Population mean $\mu = \frac{0+1}{2} = 0.5$
  - Population variance $\sigma^2 = \frac{(1-0)^2}{12} = 1/12$
  - From ISE 205 (mean and var of common distributions)
- Sampling distribution of the sample average $\bar{X}$ for $n = 10$ scored responses:
  - Mean $\mu_{\bar{X}} = \mu = 0.5$ (2.2) (regardless of n – **unbiased**)
  - Variance $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{1/12}{10} = 1/120$ (2.4) (**shrinks** as n increases)
  - Standard error $\sigma_{\bar{X}} = \sqrt{\text{Var}(\bar{X})} = \sqrt{1/120} \approx 0.091$ (2.5) (**also shrinks** in n)

Review: Computing Probabilities for Normal Distribution

Problem: Find $P(\bar{X} > 0.6)$ where $\bar{X} \sim N(0.5, 0.0913^2)$ (mean eval score, $n = 10$)



Method 1: Direct
$\bar{X} \sim N(0.5, 0.0913^2)$

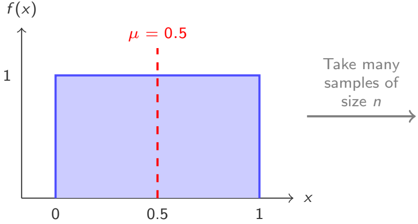
Method 2: Standardized
$Z \sim N(0, 1)$

Threshold for Z is $\frac{0.6-0.5}{0.0913} = \frac{0.1}{0.0913} \approx 1.10 \Rightarrow P(\bar{X} > 0.6) = P(Z > 1.10) \approx 0.136$

Population vs Sampling Distribution

Population Distribution

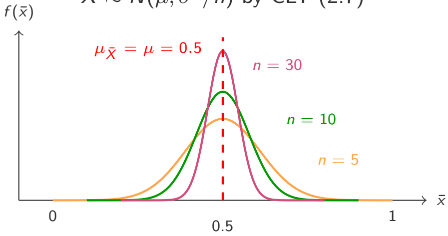
Eval scores $X_i \sim \text{Uniform}(0, 1)$



$\mu = 0.5, \quad \sigma^2 = 1/12$

Sampling Distribution of $\bar{X}$

$\bar{X} \approx N(\mu, \sigma^2/n)$ by CLT (2.7)



$\mu_{\bar{X}} = \mu$ (2.2), $\quad \sigma^2_{\bar{X}} = \sigma^2/n$ (2.4)

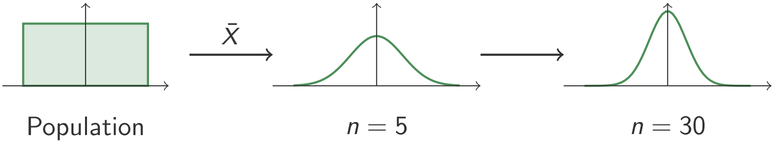
Fixed shape (doesn't change)

Shape depends on n (gets narrower as $n \uparrow$)

Both have different distributions: **same mean**, but **different shape** and **variance**

Central Limit Theorem (CLT)

Visual illustration



As $n \uparrow$, distribution of $\bar{X}$ becomes Normal and more concentrated around μ.

Central Limit Theorem (CLT)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population with mean μ and variance σ^2, and if $\bar{X}$ is the sample mean, then for large n the sampling distribution of $\bar{X}$ is approximately

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \tag{2.7}$$

Key insight: Regardless of the population distribution, $\bar{X}$ is approximately Normal for large n!

CLT for Z-statistic

Using the properties of the normal distribution, we can standardize $\bar{X}$ and calculate the Z-statistic:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n\to\infty} N(0,1) \tag{2.8}$$

For large n, the Z-statistic has a sampling distribution of the Standard Normal $N(0,1)$.

Key insight: Regardless of the population distribution, Z is approximately standard Normal for large n!

Summary #1

- The random variable $X_1, X_2, \ldots, X_n$ is a **random sample** of size n if:
  - The X_i's are **independent** random variables
  - Every X_i has the **same probability distribution**
- A statistic (also called point estimator) is a **function** of the realizations of the random variable $X_1, X_2, \ldots, X_n$ (2.1)
- A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$
- The probability distribution of a statistic is called the **sampling distribution**
- The **Central Limit Theorem (CLT)** (2.8) says the sampling distribution of $\bar{X} \approx$ Normal for large n

Questions you might be asked about sampling distribution

- Given a population distribution (say, warehouse picking times), what is the sampling distribution of the sample average $\bar{X}$ with sample size n?
- Given a population distribution, what is the probability that the sample average $\bar{X}$ is above/below a certain value?
- Given a population distribution, what is the minimum sample size n required to achieve a certain standard error for the sample average $\bar{X}$? (2.9)

Questions you might be asked about sampling distribution

Procedure 2.2

How to answer these questions? Follow Procedure 2.2:

- Find μ and σ^2 of the **population distribution**
- Use CLT to find the sampling distribution of the statistic
  - For $\bar{X}$, it will be Normal (2.7)
  - Mean: $\mu_{\bar{X}} = \mu$ (2.2) (unbiased)
  - Variance: $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$ (2.4) (shrinkage)
  - Standard error: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (2.5)
  - Sometimes, it is useful to find the standardized statistic $Z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$ (2.8)
  - Use table/calculator to find the probability using standard Normal $N(0, 1)$

Questions you might be asked about sampling distribution

What variants we might see?

- The population mean and variance might not be provided
- There are multiple populations and random samples involved
- The statistic is not the sample mean $\bar{X}$ but something else
  - Shifted mean $(\bar{X} - c)$
  - Scaled shifted mean $(a(\bar{X} - c))$—Z-statistic is an example of this
  - Sample proportion $\hat{p} = \frac{X_1+X_2+\cdots+X_n}{n}$ (2.10)
    (also a mean, but the random variables are binaries (0 or 1))
  - Difference in means $(\bar{X}_1 - \bar{X}_2)$ (2.13) or proportions $(\hat{p}_1 - \hat{p}_2)$ — Lecture 4!

Same approach:
population μ and $\sigma^2 \Rightarrow$ sampling dist $\mu_{\hat{\theta}}, \sigma^2_{\hat{\theta}} \Rightarrow$ apply $N(\mu_{\hat{\theta}}, \sigma^2_{\hat{\theta}}) \Rightarrow Z \sim N(0,1)$

What parameters we care about in this class?

- Mean μ (e.g., average container dwell time at the port)
- Variance σ^2 (e.g., spread of delivery times around the promise)
- Proportion p (e.g., guardrail false-positive rate; signup conversion rate)
- Difference in means $\mu_1 - \mu_2$ (e.g., picking time under two warehouse layouts)
- Difference in proportions $p_1 - p_2$ (e.g., conversion under two pricing pages, SAR 49 vs SAR 59)

Can you give more examples why we care about these parameters?
What estimators would you use for each parameter?

What parameters we care about in this class?

- Mean μ (population average) (sample mean $\bar{X}$)
- Variance σ^2 (spread of the data) (sample variance S^2)
- Proportion p (fraction of success in a population) (sample proportion $\hat{p}$ (2.10))
- Difference in means $\mu_1 - \mu_2$ (effect size) ($\bar{X}_1 - \bar{X}_2$)
- Difference in proportions $p_1 - p_2$ (difference in success rates) ($\hat{p}_1 - \hat{p}_2$)

Summary #2

- **General recipe** for sampling distribution problems (Procedure 2.2):
  1. Find population parameters: μ and σ^2
  2. Apply CLT to find sampling distribution: $\mu_{\hat{\theta}}$ and $\sigma^2_{\hat{\theta}}$
  3. Use Normal approximation: $\hat{\theta} \sim N(\mu_{\hat{\theta}}, \sigma^2_{\hat{\theta}})$ (2.7)
  4. (Optional) Standardize to $Z \sim N(0, 1)$ (2.8) for easier calculation
- **Quality of estimators** (Procedure 2.1): bias (2.3), variance (2.4), standard error (2.5), and MSE (2.6)
- **Remember**: Sampling distribution $\neq$ Population distribution!
  - Same mean ($\mu_{\bar{X}} = \mu$ (2.2)), but different variance ($\sigma^2_{\bar{X}} = \sigma^2/n$ (2.4))