# ISE 315: Engineering Statistics

*Lecture 4: Multi-Population Estimators*

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

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

*Course textbook: Engineering Statistics — A Handbook, Chapter 2.*

*Supplementary: Montgomery & Runger, Applied Statistics and Probability for Engineers, 6th Ed.*

# Lecture 4

Multi-Population Estimators

# Recap: Population and Sampling Distributions

- Given random sample $X_1, X_2, \ldots, X_n$ from a population with mean $\mu$
- $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ converges to a Normal Distribution (by CLT, (2.7))
  - Mean: $\mu_{\bar{X}} = \mu$ (2.2)
  - Variance:
    - If $\sigma^2$ is known: $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$ (2.4)
    - If $\sigma^2$ is unknown: $\sigma^2_{\bar{X}} \approx \frac{s^2}{n}$, where

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

## Recap: Sample Proportion

- For binary data (pass/fail, convert/leave), the parameter is a proportion $p$
- The estimator is the sample proportion (2.10): $\hat{P} = \frac{X_1+X_2+\cdots+X_n}{n}$
- Its standard error is

$$\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}} \tag{2.11}$$

- And by CLT, for large $n$, (2.12):

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

- Rule of thumb: valid when $np \geq 10$ and $n(1-p) \geq 10$

## Difference of Means $\mu_1 - \mu_2$

- Goal: estimate the difference in means between two populations with
  - $\mu_1$ and $\mu_2$ are the means
  - $\sigma_1^2$ and $\sigma_2^2$ are the variances
  - $n_1$ and $n_2$ are the sample sizes from the two populations
- We can use the difference of sample means $\bar{X}_1 - \bar{X}_2$ as an estimator
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ converges to a Normal Distribution:

$$\bar{X}_1 - \bar{X}_2 \overset{\text{approx}}{\sim} N\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \tag{2.13}$$

- with standard error

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{2.14}$$

## Similar to Single Population Estimator

- Goal: estimate unknown parameter $\Delta\mu = \mu_1 - \mu_2$
  - $\mu_1$ and $\mu_2$ are the population means
  - $\sigma_1^2$ and $\sigma_2^2$ are the population variances
  - $n_1$ and $n_2$ are the sample sizes from the two populations
- Define estimator $\Delta\hat{\mu} = \Delta\bar{X} = \bar{X}_1 - \bar{X}_2$
- The sampling distribution of $\Delta\bar{X}$ converges to a Normal Distribution (2.13):
  - Mean: $\mu_{\Delta\bar{X}} = \mathbb{E}[\Delta\bar{X}] = \Delta\mu = \mu_1 - \mu_2$
  - Variance: $\sigma^2_{\Delta\bar{X}} = \mathrm{Var}(\Delta\bar{X}) = \mathrm{Var}(\bar{X}_1) + \mathrm{Var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
  - Standard Error (2.14): $\sigma_{\Delta\bar{X}} = \sqrt{\mathrm{Var}(\Delta\bar{X})} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

# Recipe for Two-Population Probability Questions

*Procedure 2.3*

1. Write down the population parameters of **both** populations: $\mu_1, \sigma_1^2, n_1$ and $\mu_2, \sigma_2^2, n_2$ (or $p_1, p_2$)
2. Find the sampling distribution of the difference: mean and standard error ((2.13) and (2.14), or (2.15) for proportions)
3. Standardize: $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$
4. Use the standard Normal table to find the probability

*Same recipe as Procedure 2.2 — only the standard error changes.*

A logistics company compares truck fuel use under different policies. Policy A: $n_1 = 50$ trips, $\sigma_1 = 4$ L/100km. Policy B: $n_2 = 45$ trips, $\sigma_2 = 5$ L/100km. What is the probability that $\bar{X}_1 - \bar{X}_2$ differs from $\mu_1 - \mu_2$ by less than 1.5 L/100km?

*Problem 1: Truck Fuel Comparison*

---

What do we need to find first? (Procedure 2.3)

**Step 1:** Population distribution parameters: $\sigma_1 = 4$, $n_1 = 50$, $\sigma_2 = 5$, $n_2 = 45$
**Step 2:** Sampling distribution parameters (2.14):

- $\mu_{\Delta\bar{X}} = \mu_1 - \mu_2$ (unknown)
- $\sigma^2_{\Delta\bar{X}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{16}{50} + \frac{25}{45} = 0.3200 + 0.5556 = 0.8756$
- $\sigma_{\Delta\bar{X}} = \sqrt{\sigma^2_{\Delta\bar{X}}} = \sqrt{0.8756} = 0.9357$

**Step 3:** Standardize and find probability

$$P\left(|(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)| < 1.5\right) = P\left(|Z| < \frac{1.5}{0.9357}\right) = P(|Z| < 1.60)$$
$$= \Phi(1.60) - \Phi(-1.60) = 0.9452 - 0.0548 = 0.8904$$

## Difference of Sample Proportions

- Goal: estimate the difference in proportions between two populations
  - $p_1$ and $p_2$ are the true proportions of the two populations
  - $n_1$ and $n_2$ are the sample sizes
- We use $\hat{P}_1 - \hat{P}_2$ as an estimator for $p_1 - p_2$
- The sampling distribution (by CLT, for large $n_1, n_2$):

$$\hat{P}_1 - \hat{P}_2 \overset{\text{approx}}{\sim} N\left(p_1 - p_2,\ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right) \tag{2.15}$$

- Standard error: $\text{SE} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ (two copies of (2.11), added inside the root)
- Rule of thumb: Normal approximation valid when $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$

An AI safety team compares guardrail false-positive rates of two model versions on benign prompts. Version 1: $n_1 = 200$ prompts, true rate $p_1 = 0.08$. Version 2: $n_2 = 250$ prompts, true rate $p_2 = 0.05$. Find the probability that the difference in sample proportions $\hat{P}_1 - \hat{P}_2$ is within 0.03 of the true difference.

*Problem 2: Guardrail False-Positive Comparison*

---

**Step 1:** Population distribution parameters:

- $p_1 = 0.08$
- $p_2 = 0.05$
- $n_1 = 200$
- $n_2 = 250$

**Step 2:** Sampling distribution standard error (SE), from (2.15):

$$\begin{aligned}
\text{SE} &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\
&= \sqrt{\frac{0.08(0.92)}{200} + \frac{0.05(0.95)}{250}} \\
&= \sqrt{0.000368 + 0.00019} \\
&= \sqrt{0.000558} \\
&= 0.0236
\end{aligned}$$

**Step 3:** Find the probability

$$
\begin{aligned}
P\left(|(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)| \le 0.03\right) &= P\left(|Z| \le \frac{0.03}{0.0236}\right) \\
&= P(|Z| \le 1.27) \\
&= \Phi(1.27) - \Phi(-1.27) \\
&= 0.8980 - 0.1020 = \boxed{0.7960}
\end{aligned}
$$

A startup A/B tests two onboarding flows. Preliminary data suggests completion rates $p_1 \approx 0.70$ and $p_2 \approx 0.60$. How many users are needed in *each* group so that $\hat{P}_1 - \hat{P}_2$ is within 0.05 of the true difference with 95% confidence? Assume $n_1 = n_2 = n$.

*Problem 3: Sample Size for an A/B Test*

---

Same logic as the single-mean sample-size question (2.9) — new standard error.

**Step 1 — We want:** $P\left(|(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)| \le 0.05\right) = 0.95$

**Step 2 — Standard error** (with $n_1 = n_2 = n$, from (2.15)):

$$\text{SE} = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}} = \sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}}$$

**Step 3 — Condition for 95% confidence:** $\frac{0.05}{\text{SE}} = 1.96 \implies \text{SE} = \frac{0.05}{1.96} = 0.026$

**Step 4 — Solve for $n$:** with $p_1(1-p_1) + p_2(1-p_2) = 0.70(0.30) + 0.60(0.40) = 0.21 + 0.24 = 0.45$,

$$n = \frac{0.45}{(0.02551)^2} = \frac{0.45}{0.0006508} = 691.5 \implies \boxed{n = 692 \text{ users per group}}$$

## Summary: Sampling Distributions for Multi-Population Estimators

| | **Mean Difference** $(\bar{X}_1 - \bar{X}_2)$ | **Proportion Difference** $(\hat{P}_1 - \hat{P}_2)$ |
|---|---|---|
| Mean | $\mu_1 - \mu_2$ | $p_1 - p_2$ |
| Variance | $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ | $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$ |

**Remember:** For proportions, check $np \geq 10$ and $n(1-p) \geq 10$ before using Normal approximation.