# STA 103 Lecture 10: Confidence Interval (One-sample, Two-sample Inference)

Instructor: Wookyeong Song

Department of Statistics, University of California, Davis

Aug 27th, 2025

# Sampling Distribution of the One-Sample Mean

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean $\mu$ and standard deviation $\sigma$. Then,

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

- Mean: $E(\bar{X}) = E(X_1) = \mu$.
  Standard Error (SE): $\mathrm{SE}(\bar{X}) = \frac{\mathrm{sd}(X_1)}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$.

- If $X_i$ are normally distributed, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

- If $X_i$ are **not** normally distributed and $n$ is large ($n \geq 30$), then by the CLT, $\bar{X}$ is approximately normal:

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

- We assume that $\sigma^2$ is known, if not (in practice the population standard deviation $\sigma$ is rarely known), we can replace $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

# Sampling Distribution of the One-Sample Proportion

- As a special case when $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli($p$) random variables with success probability $p$. Then,

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

can be interpreted as the sample proportion of successes.
(Example: UCD students like Stats in Lecture 9, $\theta = p$, $\hat{\theta} = \hat{p}$).

- Mean: $E(\bar{X}) = E(X_1) = p$.
Standard Error (SE): $\mathrm{SE}(\bar{X}) = \frac{\mathrm{sd}(X_1)}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$.

- If $np \geq 10$ and $n(1-p) \geq 10$, then $\hat{p}$ is approximately normally distributed:

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

- These asymptotic normality form the foundation for constructing confidence intervals for means and proportions in business and economic data.

# Point Estimate vs Interval Estimate

- A point estimate provides a single value as an estimate of a population parameter.

- For example, we can use sample mean $\bar{X}$ to estimate population mean $\mu$ or sample proportion $\hat{p}$ to estimate population proportion $p$.

- However, due to sampling variability, a single point is rarely sufficient.

- Thus, we need to estimate confidence intervals, which provides a range of plausible values for the population parameter (i.e., $\mu$ or $p$) based on sample data $X_1, X_2, \ldots, X_n$.

# Confidence Interval

- **Definition**: A confidence interval (CI) for parameter $\theta$ ($\theta$ can be $\mu$ or $p$) is an interval constructed around a point estimate with a specified level of confidence, typically $95\%$, or $99\%$.

$$\begin{aligned}\theta &\in \hat{\theta} \pm \text{Z-score} \times \mathrm{SE}(\hat{\theta}) \\ &= \text{Point Estimate} \pm \text{Z-score} \times \text{SE} \\ &= \text{Best Prediction} \pm \text{Z-score} \times \text{Typical Error.}\end{aligned}$$

- **Interpretation**: A 95% confidence interval for the population mean $\mu$ means:

  *If we repeatedly drew random samples and constructed a confidence interval from each, then approximately 95% of those intervals would contain the true population mean.*

# Formulas for CI for One-Sample Mean $\mu$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. RVs with mean $\mu$

- **Case 1 (known standard deviation $\sigma$)**: The $100(1-\alpha)\%$ confidence interval for one-sample mean $\mu$ is

  $$\text{Point Estimate} \pm \text{Z-score} \times \text{SE}.$$

  $$\bar{X} \pm z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

  For $95\%$ confidence interval, i.e. $\alpha = 0.05$, then

  $$\mu \in \bar{X} \pm z_{0.975} \cdot \frac{\sigma}{\sqrt{n}}$$

- **Case 2 (unknown standard deviation $\sigma$)**: The $100(1-\alpha)\%$ confidence interval for one-sample mean $\mu$ is

  $$\text{Point Estimate} \pm \text{Z-score} \times \text{SE}.$$

  $$\bar{X} \pm z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}},$$

  where $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$. (Student's t-distribution could be used instead, but will not get into the detail)

# Formulas for CI for One-Sample Proportion $p$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli($p$) RVs with success probability $p$.

- **Case 1 (plug-in estimator in Lec 9)**: The $100(1-\alpha)\%$ confidence interval for one-sample proportion $p$ is

  Point Estimate $\pm$ Z-score $\times$ SE.

  $$\hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- **Case 2 (conservative approach in Lec 9)**: The $100(1-\alpha)\%$ confidence interval for one-sample proportion $p$ is

  Point Estimate $\pm$ Z-score $\times$ SE.

  $$\hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{0.5(1-0.5)}{n}}$$

# Example for CI for One-Sample Mean $\mu$

- **Example**: An operations manager at an e-commerce company wants to estimate the **average delivery time** (in days) for standard shipping within California. Accurate estimation for delivery time is key to customer satisfaction, and management wants to report the average delivery time with a $95\%$ confidence interval.

- **Data Collection**: The manager randomly selects a simple random sample of $n = 40$ recent orders and records the delivery times. The summary statistics from the sample are:
  - Sample mean: $\bar{X} = 2.6$ days
  - Sample standard deviation: $s = 0.8$ days
  - Sample size $n = 40$

# Example for CI for One-Sample Mean $\mu$

- Even though $X_i$'s are not normally distributed, $n$ is large ($n \geq 30$).

- We do not know the population standard deviation $\sigma$, so we need to use sample standard deviation $s$ instead.

- **Answer**: The $95\%$ confidence interval, we have $\alpha = 0.05$, then

$$\begin{aligned}\bar{X} \pm z_{0.975} \times \mathrm{SE}(\bar{X}) &= \bar{X} \pm z_{0.975} \times \frac{s}{\sqrt{n}} \\ &= 2.6 \pm 1.96 \times \frac{0.8}{\sqrt{40}} = [2.466, 2.734].\end{aligned}$$

- **Interpretation**: The operations manager can be $95\%$ confident that the average delivery time for all standard California orders is between 2.466 and 2.734 days.

- If the company promised "delivery in under 3 days," this interval **supports** the claim.

# Example for CI for One-Sample Proportion $p$

- **Example (Lec 9)**: Estimating a population proportion $p$ of all UCD students who like statistics. Interview random $1000$ UCD students. $34\%$ said they like stats. Find $99\%$ confidence interval for $p$.

- **Statistical Thinking**: Let $X_i$ be whether $i$th student like stats or not, (i.e., $X_i = 1$ if $i$th student likes stats, and $X_i = 0$ if does not). Then, we observe $X_1, X_2, \ldots, X_{1000}$. One may use the box model.

- Here, $p = \theta =$ proportion of students who like stats over all students. Our estimator $\hat{p} = \hat{\theta} =$ proportion of students who like stats in the $1000$ samples (data).

$$\hat{p} = \hat{\theta} = \frac{X_1 + X_2 + \cdots X_{1000}}{1000} = \bar{X} = 0.34.$$

# Example for CI for One-Sample Proportion $p$

The $100(1-\alpha)\% = 99\%$ confidence interval corresponds to $\alpha = 0.01$. Then $z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.01}{2}} = z_{0.995}$.

- **Case 1 (plug-in estimator in Lec 9)**: The $99\%$ confidence interval for one-sample proportion $p$ is

$$\hat{p} \pm z_{0.995} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.34 \pm 2.58 \times \sqrt{\frac{0.34 \times 0.66}{1000}}.$$

- **Case 2 (conservative approach in Lec 9)**: The $99\%$ confidence interval for one-sample proportion $p$ is

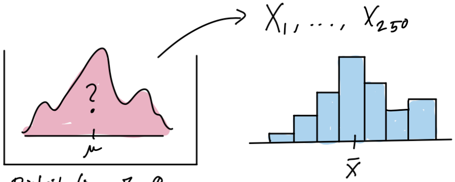
$$\hat{p} \pm z_{0.995} \cdot \sqrt{\frac{0.5(1-0.5)}{n}} = 0.34 \pm 2.58 \times \sqrt{\frac{0.5 \times 0.5}{1000}}.$$

# Comparing Two Populations

- The basic reasoning used in the One-Sample mean $\mu$ or proportion $p$ works in more complicated settings.
- Here is an example that tests the difference between two treatments.
- **Example 1**: Suppose you have developed a new drug for lowering cholesterol and want to test if it is effective.
- You get a random sample of $250$ people and give them the drug, then measure $X =$ "the % reduction of cholesterol after 6 months." Let $X_1, X_2, \ldots, X_{250}$ denote the $X$ observations for each patient.

# Comparing Two Populations

- You get a random sample of $250$ people and give them the crug, then measure $X =$ "the % reduction of cholesterol after 6 months." Let $X_1, X_2, \ldots, X_{250}$ denote the $X$ observations for each patient.




Distribution of % reduction in cholesterol if the new drug is given to everyone.

$\mu = ?$

$\sigma = ?$

The % reduction in the sample

$\bar{X} = 7.2\%$

$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = 4.1\%$

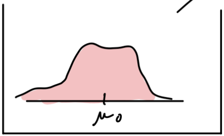
- This might suggest $\mu > 0$ but we need to rule out a placebo effect.

# Comparing Two Populations

- To account for the placebo effect, you sample $500$ samples (called the control group) and give them a sugar pill.

- Let $Y =$ "% reduction of cholesterol after 6 months," and $Y_1, Y_2, \ldots, Y_{500}$ denote the measurements for the patients in the control group.

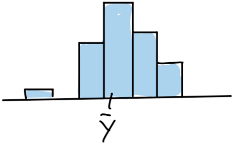
$\rightarrow Y_1, Y_2, \ldots, Y_{500}$




Distribution of % reduction in cholesterol when taking a sugar pill

$\mu_0 = ?$

$\sigma_0 = ?$




The % reduction in the sample

$\bar{Y} = 3.9\%$

$\tilde{\sigma}_0 = \sqrt{\frac{1}{500}\sum_{i=1}^{500}(Y_i - \bar{Y})^2} = 2.5\%$

# Two-sample Inference

- **Question 1**: Find $95\%$ Confidence Interval for $\mu - \mu_0$?

- **Answer**: Recall that

$$\begin{aligned}\theta &\in \hat{\theta} \pm \text{Z-score} \times \mathrm{SE}(\hat{\theta}) \\ &= \text{Point Estimate} \pm \text{Z-score} \times \text{SE} \\ &= \text{Best Prediction} \pm \text{Z-score} \times \text{Typical Error.}\end{aligned}$$

- $\theta = \mu - \mu_0$.

- The point estimate is $\hat{\theta} = \bar{X} - \bar{Y}$. Then,

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu - \mu_0.$$

- Standard Error (SE) is $\mathrm{SE}(\hat{\theta}) = \mathrm{SE}(\bar{X} - \bar{Y})$.

- $95\%$ Z-score $\rightarrow \alpha = 0.05 \rightarrow z_{1-\frac{0.05}{2}} = z_{0.975} = 1.96$.

# Two-sample Inference

- **Answer (Conti.)** Since $X$'s are independent of the $Y$'s,

$$\begin{aligned}\mathrm{SE}(\bar{X}-\bar{Y}) = \mathrm{sd}(\bar{X}-\bar{Y}) &= \sqrt{\mathrm{Var}(\bar{X}-\bar{Y})} \\ &= \sqrt{\mathrm{Var}(\bar{X})+\mathrm{Var}(\bar{Y})} = \sqrt{\frac{\sigma^2}{250}+\frac{\sigma_0^2}{500}} \approx \sqrt{\frac{\hat{\sigma}^2}{250}+\frac{\hat{\sigma}_0^2}{500}} \\ &= \sqrt{\frac{4.1^2}{250}+\frac{2.5^2}{500}} = \sqrt{0.0797} = 0.282.\end{aligned}$$

- The typical error when using $\bar{X}-\bar{Y}$ to estimate $\mu-\mu_0$ is about $\mathrm{SE}(\bar{X}-\bar{Y}) = 0.282$. Then, approximately $95\%$ of the time,

$$\bar{X}-\bar{Y} \approx \mu-\mu_0 \pm 1.96 \times 0.282.$$

- Moving things around,

$$\begin{aligned}\mu-\mu_0 \in \bar{X}-\bar{Y} \pm z_{0.975} \times \mathrm{SE}(\bar{X}-\bar{Y}) &= \bar{X}-\bar{Y} \pm 1.96 \times 0.282 \\ &= (7.2-3.9) \pm 1.96 \times 0.282 = (2.747, 3.853),\end{aligned}$$

with $95\%$ confidence. (**Interpretation**: most likely $\mu > \mu_0$, the drug works better than a placebo, since both the lower bound and upper bound are positive.)

# Two-sample Inference

- **Question 2** Quantify the amount of evidenced that $\mu - \mu_0 > 0$ from the data.

- **Answer**: if it was actually the case that $\mu - \mu_0 \leq 0$, (i.e. $\mu \leq \mu_0$) then we just observed $\bar{X} - \bar{Y}$ to have a Z-score greater than $11.74$ since if $\mu - \mu_0 \leq 0$, then

$$Z = \frac{\bar{X} - \bar{Y} - (\mu - \mu_0)}{\sqrt{\frac{\sigma^2}{250} + \frac{\sigma_0^2}{500}}} \geq \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{\sigma^2}{250} + \frac{\sigma_0^2}{500}}}$$

$$\approx \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{\hat{\sigma}^2}{250} + \frac{\hat{\sigma}_0^2}{500}}} = \frac{3.3}{0.282} = 11.74.$$

- The probability of that happening is

$$P(Z > 11.74) < 10^{-10}.$$

  Later, this results $P(Z > 11.74)$ is called the p-value for theting the null hypothesis $\mu - \mu_0 \leq 0$.

- The data gives conclusive evidence that $\mu - \mu_0 > 0$.

# Two-sample Inference

- **Example 2**: At a large university you take a random sample of in-state students, $X_1, X_2, \ldots, X_{n_1}$ and out-of-state students $Y_1, Y_2, \ldots, Y_{n_2}$. Assume that $X_i$'s and $Y_i$'s are independent. Find the following data on their GPAs:

| in-state | out-of-state |
|---|---|
| $\bar{X} = 2.8$ | $\bar{Y} = 3.0$ |
| $s_X = 0.4$ | $s_Y = 0.5$ |
| $n_1 = 25$ | $n_2 = 29$ |

- Let

$$\mu_X = E(X_1) = \text{Average GPA for all in-state students}$$
$$\mu_Y = E(Y_1) = \text{Average GPA for all out-of-state students}$$
$$\sigma_X^2 = \mathrm{Var}(X_1) = \text{Variance of GPA for all in-state students}$$
$$\sigma_Y^2 = \mathrm{Var}(Y_1) = \text{Variance GPA for all out-of-state students}$$

- **Question**: Build an approximately $99.7\%$ CI for $\mu_X - \mu_Y$.

# Two-sample Inference

- **Answer**: We can estimate $\mu_X - \mu_Y$ by $\bar{X} - \bar{Y}$ and we have

$$\bar{X} - \bar{Y} \in (\mu_X - \mu_Y) \pm z_{1-\frac{\alpha}{2}} \times \mathrm{SE}(\bar{X} - \bar{Y}).$$

- Moving things around, we have

$$\mu_X - \mu_Y \in (\bar{X} - \bar{Y}) \pm z_{1-\frac{\alpha}{2}} \times \mathrm{SE}(\bar{X} - \bar{Y}).$$

- Notice that $\bar{X} - \bar{Y} = 2.8 - 3.0 = -0.2$.
- Here $99.7\%$ CI $\rightarrow \alpha = 0.003$, then $z_{1-\frac{0.003}{2}} = z_{0.9985} = 3$.
- Also notice that

$$\mathrm{SE}(\bar{X} - \bar{Y}) = \mathrm{sd}(\bar{X} - \bar{Y}) = \sqrt{\mathrm{Var}(\bar{X} - \bar{Y})} = \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}$$

$$\approx \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}} = \sqrt{\frac{0.4^2}{25} + \frac{0.5^2}{29}} = 0.12.$$

# Two-sample Inference

- **Answer (Conti.)**: An approximate $99.7\%$ CI for $\mu_X - \mu_Y$ is

$$\begin{aligned} \mu_X - \mu_Y &\in (\bar{X} - \bar{Y}) \pm 3 \times \mathrm{SE}(\bar{X} - \bar{Y}) \\ &= -0.2 \pm 3 \times (0.12) = [-0.56, 0.12], \end{aligned}$$

  which suggests there is not enough to decide if $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$. (This is because the CI contains $0$.)