Lecture 9: Confidence Intervals

Module 3: part 2

Spring 2025

Logistics

- Solutions of Assessment for Module 1 is available online
- Lab next week about simulations and sampling distributions

Linear Model Fundamentals

- Linear Model Framework:

$$Y_i = b_0 + \sum_j b_j X_{ij} + \varepsilon_i$$

- Key Properties:
  - $E(\hat{b} \mid X) = b$ (Unbiased estimation)
  - For simple linear regression:

$$\text{var}(\hat{b}_1 \mid \mathbf{X}) = \frac{\sigma_\varepsilon^2}{(n-1)s_x^2}$$

- Variance decreases with:
  - Larger sample size ($n \uparrow$)
  - Lower error variance ($\sigma_\varepsilon^2 \downarrow$)
  - More spread out covariates ($s_x^2 \uparrow$)

Understanding Errors and Residuals

- Error Variance Estimation:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2$$

- Residual Decomposition:

$$y_i - \hat{y}_i = (b_0 - \hat{b}_0) + \sum_j (b_j - \hat{b}_j)x_{ij} + \varepsilon_i$$

- Key Distinctions:
  - Residual: Observable difference $(y_i - \hat{y}_i)$
  - Error: Unobservable component (ε_i)

Multiple Linear Regression: Variance Structure

- Full Variance-Covariance Matrix:

$$\text{var}(\hat{\mathbf{b}} \mid X) = \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- Matrix Structure:

$$\begin{bmatrix} \text{var}(\hat{b}_0) & \text{cov}(\hat{b}_0, \hat{b}_1) & \cdots \\ \text{cov}(\hat{b}_0, \hat{b}_1) & \text{var}(\hat{b}_1) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

- Error Variance Estimation:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n - (p+1)} RSS(\hat{\mathbf{b}})$$

Impact of Collinearity

- For 2 standardized covariates (mean = 0, variance = 1):

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

- Resulting variance structure:

$$\text{var}(\hat{\mathbf{b}} \mid X) = \sigma_\varepsilon^2 \begin{bmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{bmatrix}$$

- Key Implications:
  - As $|\rho| \to 1$, variance $\to \infty$
  - Estimates remain unbiased but precision decreases
  - Similar predictions despite different coefficients

Confidence Intervals

Confidence Intervals: A Motivating Example

- Question: Given our observed data, what is a plausible range for a parameter?
- Case Study: House Price Model
  - Model includes interaction between quality and age
  - Estimated age coefficient: $\hat{\beta}_{age} = -0.0046$
  - Key Questions:
    - Could the true coefficient be 0?
    - Could the true coefficient be positive?

Understanding Confidence Intervals

Definition

A confidence interval is a **procedure** that produces intervals which, when applied to **new data**, will contain the true parameter a fixed proportion of the time (e.g., 95

- Critical Distinctions:
  - **Correct**: "This procedure, when used repeatedly, produces intervals containing the true parameter 95% of the time".
  - **Incorrect**: "There is a 95% chance that this interval (1.5, 2.6) contains the true parameter".
- Key Insights:
  - Once computed, an interval either contains the parameter or does not.
  - The probability relates to the **procedure**, not to the specific interval.

Analogy

Consider $X \sim N(0,1)$:

- $P(X > 0) = 0.5$ before drawing X
- If we observe $X = 2.5$, $P(2.5 > 0) = 1$
- The probability applies to the process, not the outcome.

Standardizing Coefficient Estimates

Known Variance Case

When $\text{var}(\hat{b}_k)$ is known:

$$\frac{\hat{b}_k - b_k}{\sqrt{\text{var}(\hat{b}_k)}} \sim N(0, 1)$$

Estimated Variance Case

When variance is estimated (using standard error):

$$\frac{\hat{b}_k - b_k}{\sqrt{\widehat{\text{var}}(\hat{b}_k)}} \sim T_{n-p-1}$$

Standardizing Coefficient Estimates

- Key Terms:
  - $\widehat{\text{var}}(\hat{b}_k)$ is called the **standard error**
  - $n - p - 1$ represents degrees of freedom
    - $n =$ number of observations
    - $p =$ number of predictors
- Important Note:
  - T distribution has heavier tails than Normal
  - As $n \to \infty$, T distribution approaches Normal

T distributions

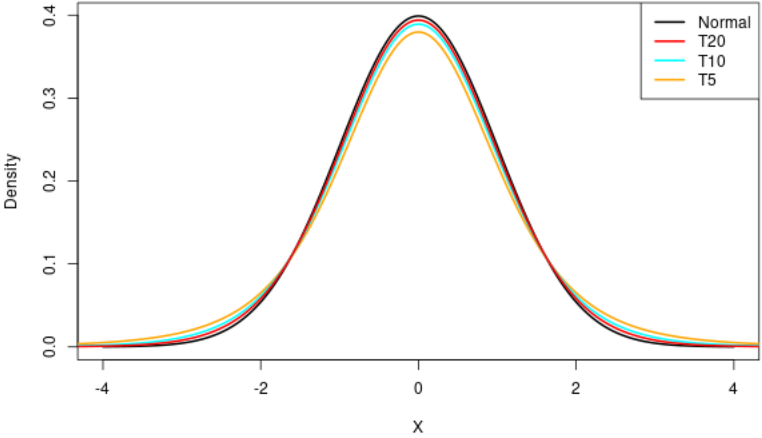


Figure: Comparing T distributions to a normal

T distributions

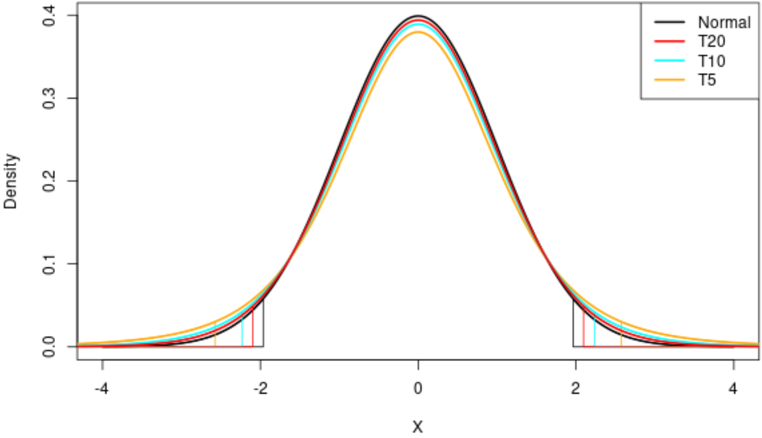
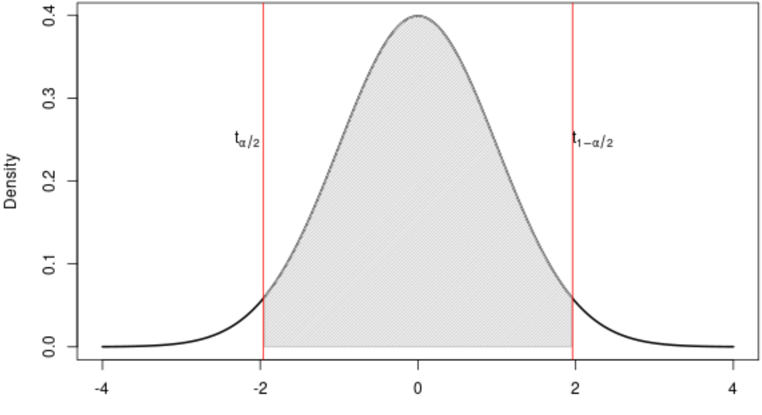


Figure: Comparing T distributions to a normal

Confidence Interval for single coefficient

If t is drawn from a T_{n-p-1} then we can look up values such that for $0 < \alpha < 1$, we have

$$P(t_{\alpha/2,n-p-1} < t < t_{1-\alpha/2,n-p-1}) = 1 - \alpha$$



Confidence Interval for single coefficient

If t is drawn from a T_{n-p-1} then we can look up values such that for $0 < \alpha < 1$, we have

$$\begin{aligned}
1-\alpha &= P\left(t_{\alpha/2,n-p-1} < \frac{\hat{b}_1 - b_1}{\sqrt{\widehat{\text{var}}(\hat{b}_1)}} < t_{1-\alpha/2,n-p-1}\right) \\
&= P\left(t_{\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)} < \hat{b}_1 - b_1 < t_{1-\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)}\right) \\
&= P\left(t_{\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)} - \hat{b}_1 < -b_1 < t_{1-\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)} - \hat{b}_1\right) \\
&= P\left(\hat{b}_1 - t_{1-\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)} > b_1 > \hat{b}_1 - t_{1-\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)}\right)
\end{aligned}$$

Understanding Confidence Intervals

General Form

For a single coefficient, the interval is:

$$\left(\hat{b}_1 - t_{\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)}, \hat{b}_1 - t_{\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)}\right)$$

Simplified Form

Since $t_{\alpha/2,n-p-1} = -t_{1-\alpha/2,n-p-1}$, we can write:

$$\hat{b}_1 \pm t_{1-\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)}$$

- Components:
  1. **Estimate** ($\hat{b}_1$): Point estimate of coefficient
  2. **Multiplier** ($t_{1-\alpha/2}$): Based on confidence level
  3. **Standard Error** ($\sqrt{\widehat{\text{var}}(\hat{b}_1)}$): Estimated variability

Understanding Confidence Intervals

General Form

For a single coefficient, the interval is:

$$\left(\hat{b}_1 - t_{\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)}, \hat{b}_1 - t_{\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)}\right)$$

Simplified Form

Since $t_{\alpha/2,n-p-1} = -t_{1-\alpha/2,n-p-1}$, we can write:

$$\hat{b}_1 \pm t_{1-\alpha/2,n-p-1}\sqrt{\widehat{\text{var}}(\hat{b}_1)}$$

Correct Interpretation

We are $(1-\alpha)\%$ confident that the true parameter lies in this interval.

Practical Exercise: Computing Confidence Intervals

Model Setup

$$Y_i = b_0 + b_1 X_1 + \varepsilon_i, \quad i = 1, \ldots, 100$$

- Critical Values:
  - 95% CI: $t_{.025,98} = 1.984$
  - 90% CI: $t_{.05,98} = 1.661$
- Tasks:
  1. Verify your 95% CI matches the formula:

$$\hat{b}_1 \pm t_{.025,98} \sqrt{\widehat{\text{var}}(\hat{b}_1)}$$

  2. Calculate 90% CI using:

$$\hat{b}_1 \pm t_{.05,98} \sqrt{\widehat{\text{var}}(\hat{b}_1)}$$

Remember

A confidence interval is a **procedure** that produces intervals containing the true parameter $(1-\alpha)\%$ of the time when applied to **new data**.

Confidence Intervals for Conditional Mean

Goal

Estimate plausible values for $E(Y_i \mid X = x) = b_0 + b_1 x$

- Example Interpretation:
  - "We are 95% confident that the average price of a home with 1000 sq ft is between $L and $U"
- Formula:

$$\hat{b}_0 + \hat{b}_1 x \pm t_{\alpha/2, n-(p+1)} \times \text{SE}$$

- Standard Error:

$$\text{SE} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Key Insights

- SE increases as x moves away from $\bar{x}$
- SE decreases with larger sample size
- Formula extends to multiple regression (computed by software)
- **Key Assumption**: Errors (ε) are normally distributed

Questions

- All things equal, what will typically be wider? A 95% confidence interval or a 90% confidence interval?
- All things equal, what will typically be wider? A 95% confidence interval when $n = 100$ or when $n = 500$ where n is the number of observations?
- All things equal, what will typically be wider? A 95% confidence interval when $p = 5$ or when $p = 10$ where p is the number of predictors in the model?

Wrap up

- Confidence intervals reflect uncertainty we have in estimating parameters
- Can form confidence interval for regression parameters
- Can form confidence interval for conditional mean parameters
- Can form prediction interval for individual observations