Lecture Module 5 - Economics 113

- Agenda
  1. Sampling Distribution
  2. T-statistic
  3. T-Tests - One-sided and Two-sided
  4. Confidence Intervals
  5. P-Values

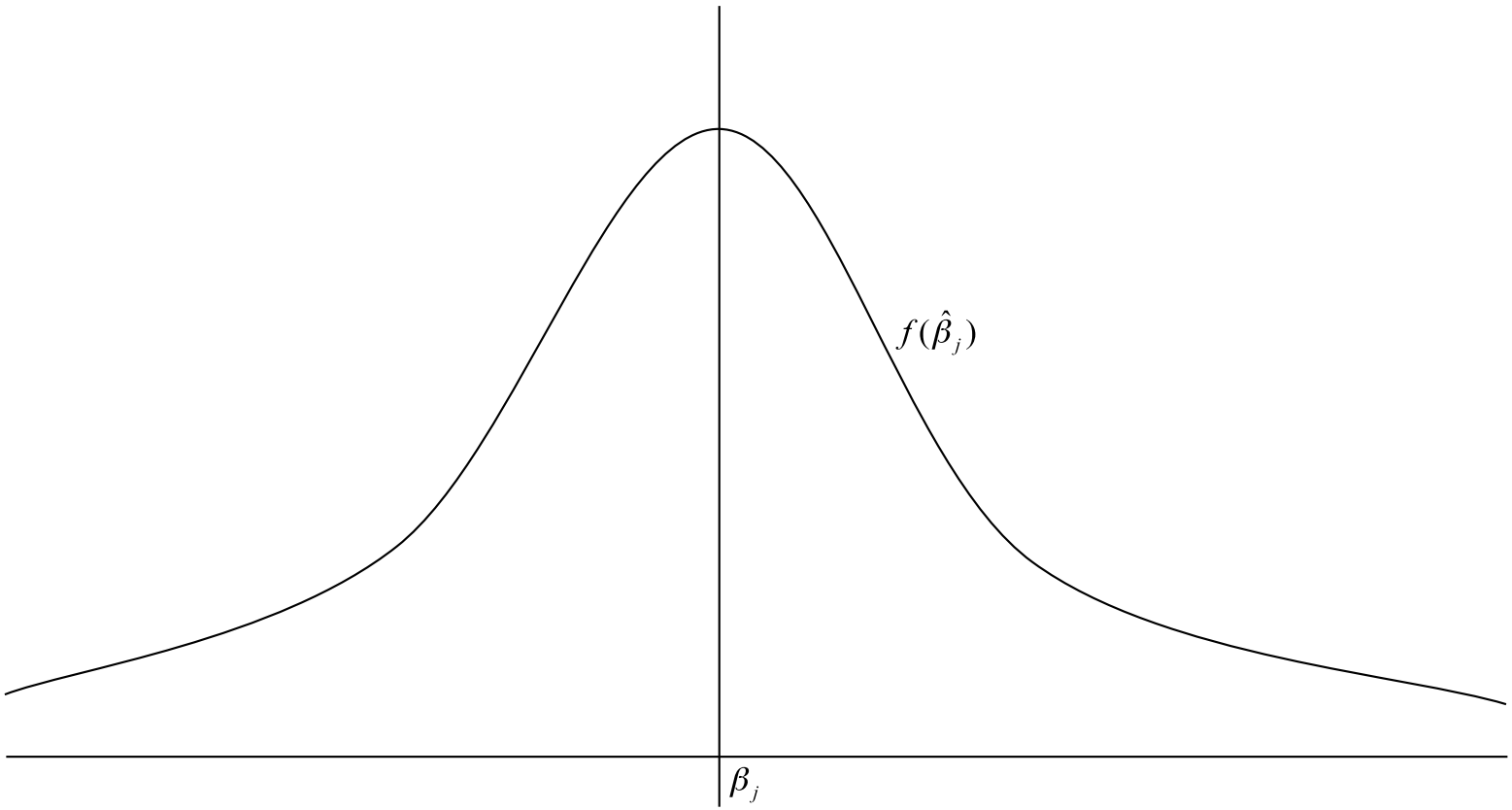
Multivariate Regression

- If the following assumptions hold:
  1. Linear in parameters β_0, β_1,...,β_k
  2. Random Sampling
  3. Zero conditional mean
  4. No perfect collinearity
  5. Homoskedasticity
- OLS is the Best Linear Unbiased Estimator (OLS is BLUE)
- Assumptions 1-5 are referred to as the "Gauss-Markov" assumptions

Multivariate Regression

- A sixth assumption: $u \tilde{} Normal(0, \sigma^2)$
- With this assumption:

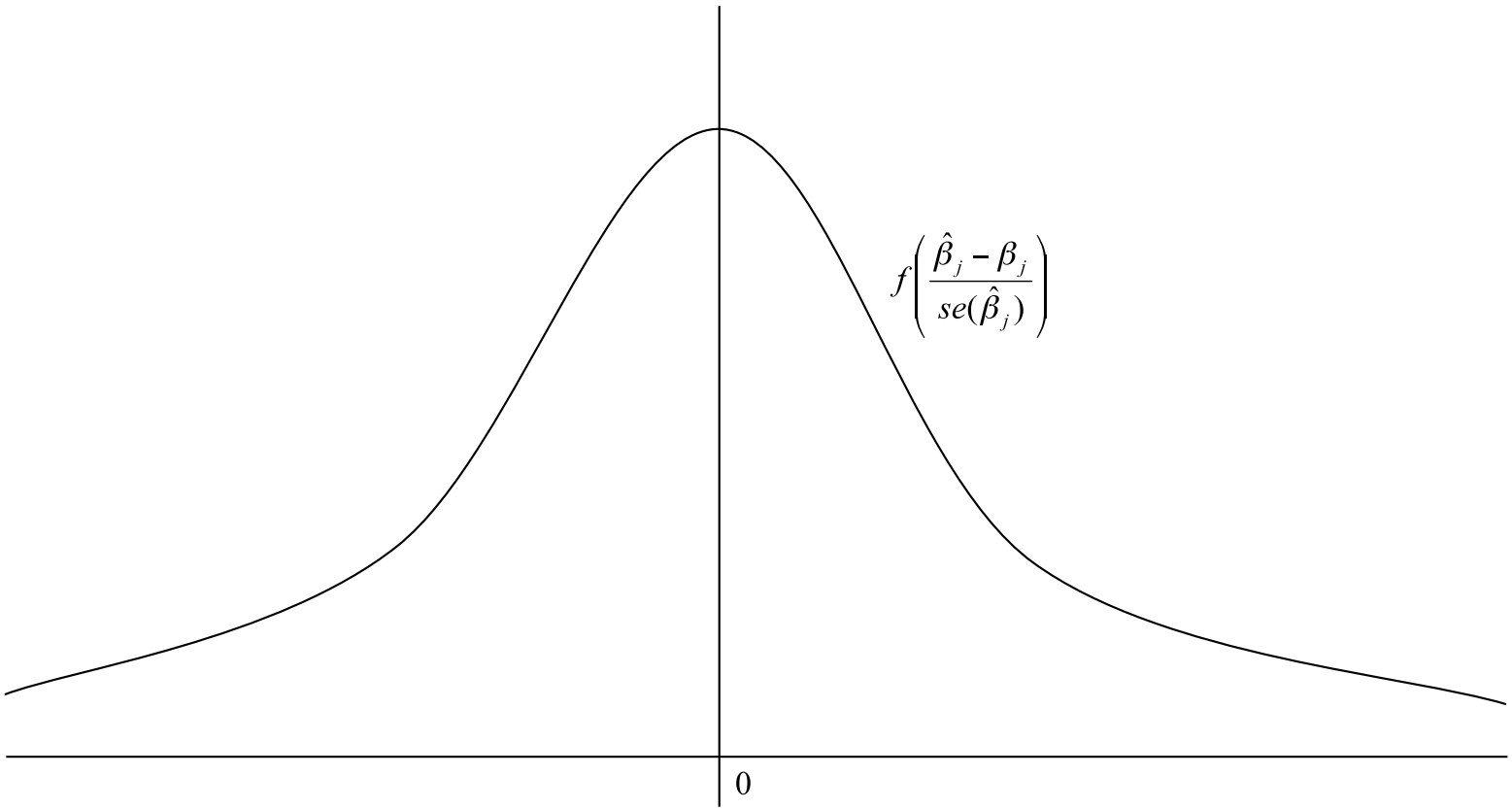
$$\widehat{\beta}_j \tilde{} Normal(\beta_j, Var(\widehat{\beta}_j))$$


$f(\hat{\beta}_j)$
β_j

Multivariate Regression

Normalization

- If the $\widehat{\beta}_j$'s are normal, what can we do with them?
- Normalize them! $\frac{\widehat{\beta}_j - \beta_j}{sd(\widehat{\beta}_j)} \tilde{} Normal(0,1)$
- Distribution of $\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)}$ is standard normal distribution if N is large.


$f\left(\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}\right)$
0

Multivariate Regression

The T-Distribution

- More generally:

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \sim t_{n-k-1}$$

- t_{n-k-1} represents the "t-distribution" with $n-k-1$ degrees of freedom.
  - k slope parameters, 1 intercept term, n observations
- $se(\widehat{\beta}_j)$ is the standard error of $\widehat{\beta}_j$

$$se(\widehat{\beta}_j) = \sqrt{Var(\widehat{\beta}_j)}$$

- t_{n-k-1} looks similar to a standard normal distribution.
- This is the 'correct' distribution of $\widehat{\beta}_j's$, centered around the population value β_j

The t-statistic

The breakdown

$$t(\widehat{\beta}_j|\beta_j) = \frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)}$$

- $\widehat{\beta}_j$ and $se(\widehat{\beta}_j)$ are estimates
- β_j is the population parameter

  → We can't measure it, so we **form a hypothesis** about it

- If $H_0 : \beta_j = 0$

$$t(\widehat{\beta}_j|\beta_j = 0) = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

- If $H_0 : \beta_j = \beta_H$

$$t(\widehat{\beta}_j|\beta_j = \beta_H) = \frac{\widehat{\beta}_j - \beta_H}{se(\widehat{\beta}_j)}$$

The t-statistic

The breakdown

- The farther $\widehat{\beta}_j$ is from the hypothesized value, the more likely the hypothesis is incorrect.
- $H_0 : \beta_j = 0$
- If H_0 is true, $\widehat{\beta}_j$ should be close to 0.

$$\Rightarrow \quad \text{small} \left| \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \right|$$

- if H_0 is false, $\widehat{\beta}_j$ should be far away from 0

$$\Rightarrow \quad \text{large} \left| \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \right|$$

- What defines large and small?
- T-distribution defines large and small.

Hypothesis Testing

- "Null Hypothesis": The hypothesis that we are testing.
- "Alternative Hypothesis": Hypothesis which we are testing the null against.
- Two outcomes:
  - Reject the null in favor of an alternative.
  - Fail to reject the null.
- We tend to reject the null when the estimate is "far away" from the null, toward the alternative hypothesis.
- Given an estimate and standard error (precision), "far away" is determined by the t-distribution.
- For the moment, we are testing the following:

$$H_0 : \beta_j = 0$$

- Why will we never accept this hypothesis as the truth?

The t-statistic

one sided test

- Experience and Wages

$$\log(wage) = \beta_0 + \beta_1 Educ + \beta_2 Exper + \beta_3 Tenure + u$$

- We estimate:

$$\begin{aligned} \widehat{\log(wage)} &= \underset{(0.104)}{0.284} + \underset{(0.007)}{0.092} Educ + \underset{(0.0017)}{0.0041} Exper + \underset{(0.003)}{0.022} Tenure \\ obs &= 526, R^2 = 0.316 \end{aligned}$$

$$\underset{\text{(standard error)}}{estimate}$$

- Suppose $H_0 : \beta_2 = 0$.
- T-statistic for this hypothesis?
- Use the formula

$$t(\widehat{\beta}_2 | \beta_2 = 0) = \frac{\widehat{\beta}_2 - 0}{se(\widehat{\beta}_2)} = \frac{0.0041}{0.0017} = 2.41$$

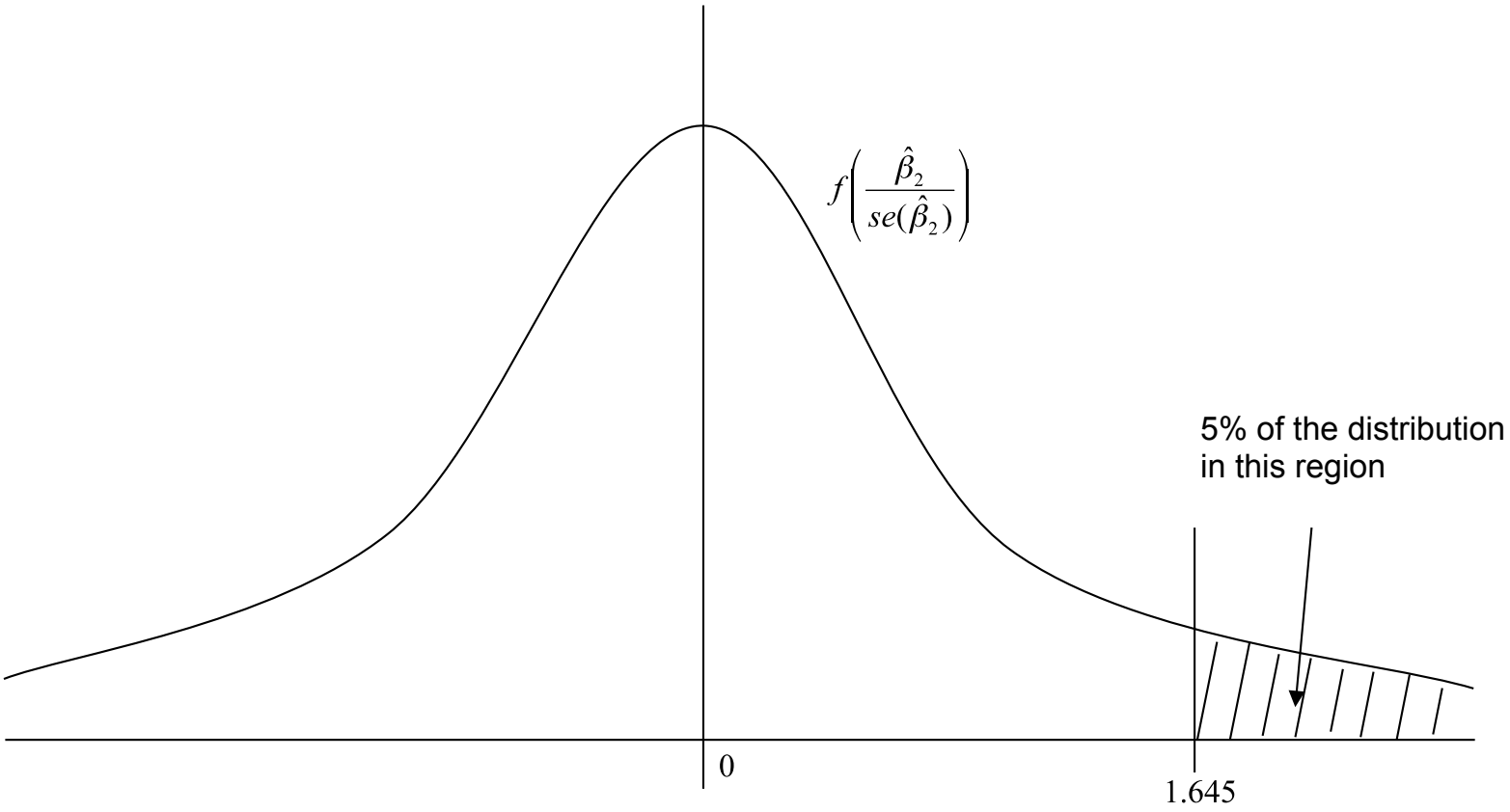
The t-statistic

one sided test

- We need to compare our null hypothesis to some alternative, H_A.
- $H_A : \beta_2 > 0$
- When should we reject H_0 in favor of H_A?

  $\Rightarrow$ If $\widehat{\beta}_2$ is sufficiently greater than zero.
- What determines "sufficiently greater"

  $\Rightarrow$ The t-distribution
- However, we could be wrong, since we do not measure β_2
- 95% "Confidence level" is a common desired level of accuracy.
- This means that we falsely reject the null no more than 5% of the time.


$f\left(\frac{\hat{\beta}_2}{se(\hat{\beta}_2)}\right)$
5% of the distribution
in this region
0
1.645

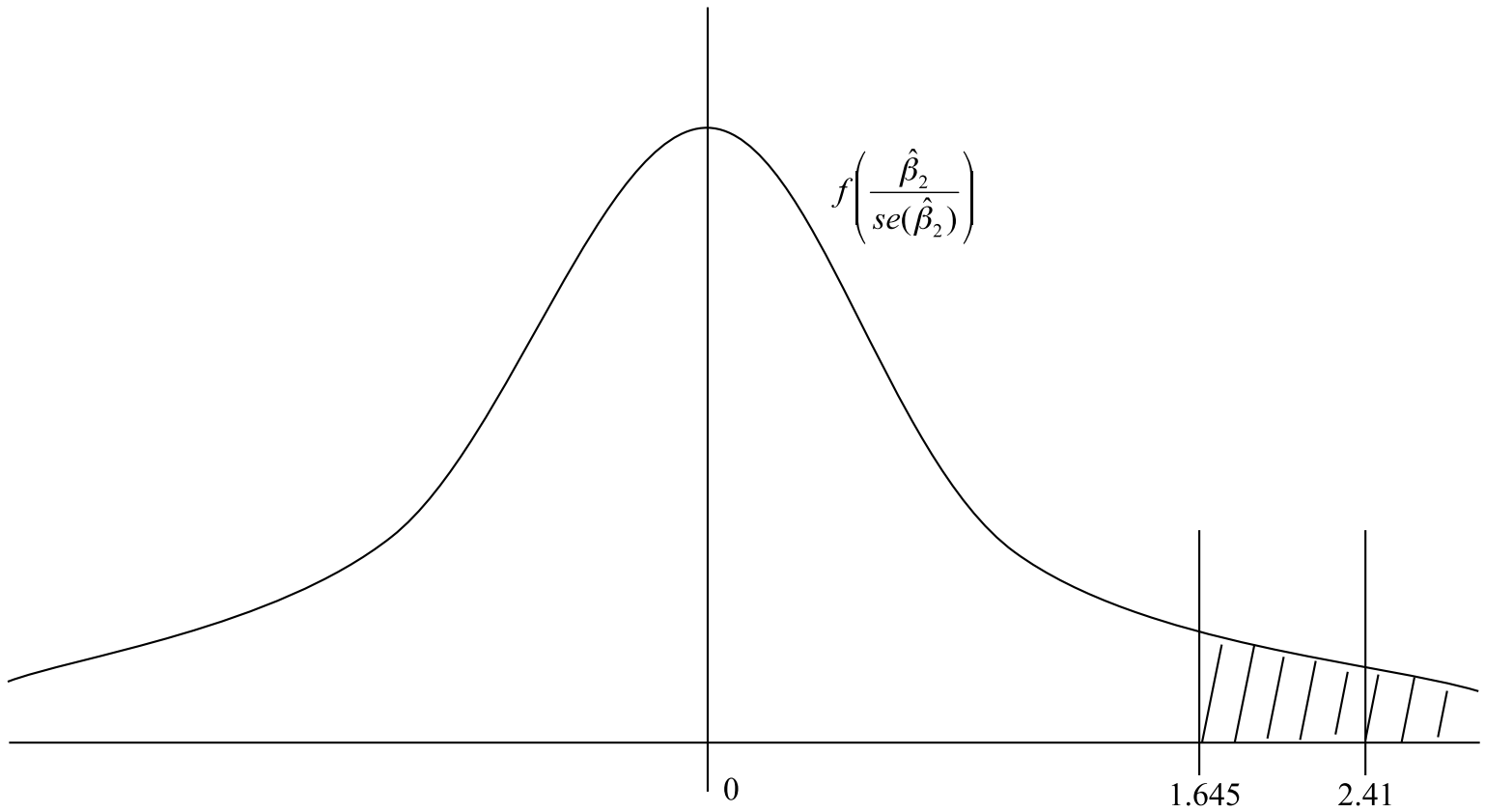
The t-statistic

one sided test

- $H_A : \beta_2 > 0$ - 95% confidence
- Three steps
  1. Calculate $n - k - 1$
  2. If $n - k - 1$ large enough, find the appropriate critical value, t_{crit}, using the standard normal distribution.
  3. If $\frac{\widehat{\beta}_2 - 0}{se(\widehat{\beta}_2)} > t_{crit}$, reject H_0 in favor of H_A.

     $\Rightarrow$ Thus, if $\widehat{\beta}_2$ is sufficiently *greater* than zero, it is very unlikely that $\beta_2 = 0$.

     $\Rightarrow$ It would require a fluke for $\widehat{\beta}_2$ to be significantly greater than its true value.
- With enough observations, at 95% confidence, $t_{crit} = 1.645$
- Since $t(\widehat{\beta}_2 | \beta_2 = 0) = 2.41 > 1.645$, **reject** $\beta_2 = 0$ in favor of $\beta_2 > 0$.
- Experience has a **positive and statistically significant** effect on wages!


$f\left(\frac{\hat{\beta}_2}{se(\hat{\beta}_2)}\right)$
0
1.645
2.41

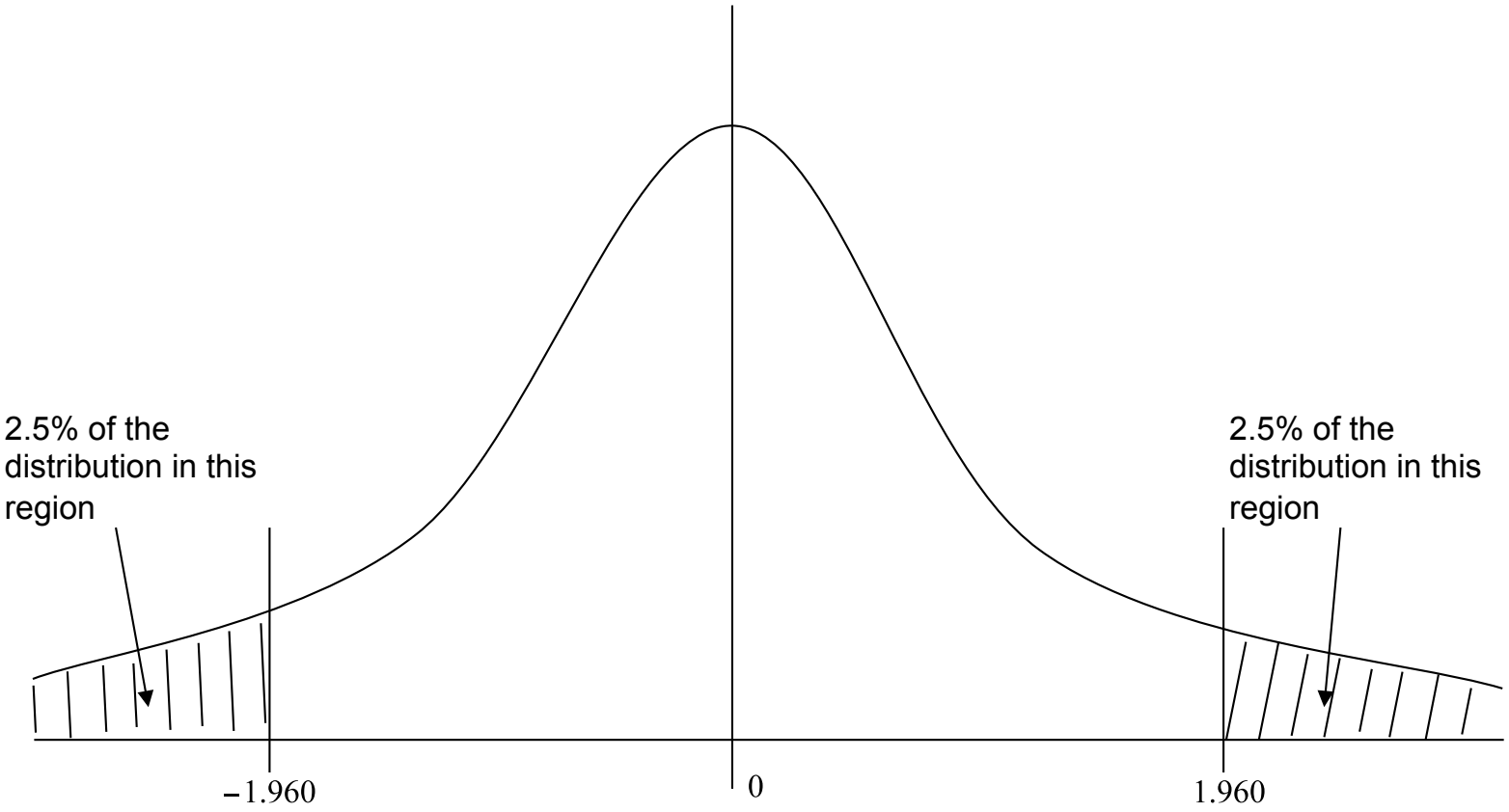
The t-statistic

two-sided test

- Two-sided tests require a slightly different approach.
- Now, the alternative is $H_A : \beta_2 \neq 0$
- Three steps
  1. Calculate $n - k - 1$
  2. If $n - k - 1$ large enough, find the two-sided critical value, t_{crit}, using the standard normal distribution.
  3. If $\left| t(\widehat{\beta}_2 | \beta_2 = 0) \right| > t_{crit}$, reject H_0 in favor of H_A.

$\Rightarrow$ Thus, if $\widehat{\beta}_2$ is sufficiently far away from zero, it is very unlikely that $\beta_2 = 0$.

$\Rightarrow$ $\widehat{\beta}_2$ can now either be positive or negative.


2.5% of the distribution in this region
2.5% of the distribution in this region
–1.960
0
1.960

The t-statistic

two-sided test

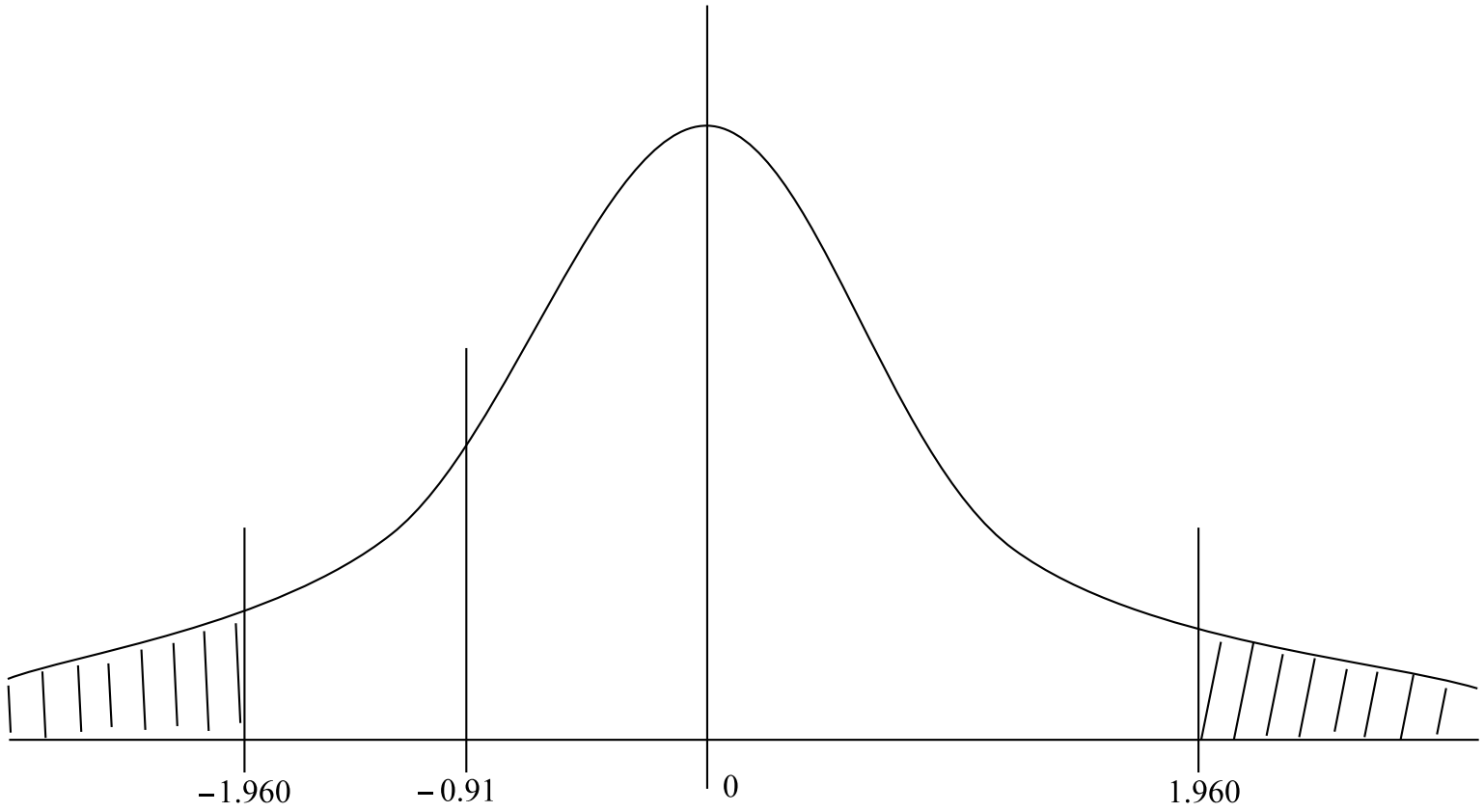
- Test scores and school size. We estimate:

$$\begin{aligned} \widehat{math10} &= \underset{(6.113)}{2.274} + \underset{(0.00010)}{0.00046}totcomp + \underset{(0.040)}{0.048}staff - \underset{(0.00022)}{0.0002}enroll \\ obs &= 408,\ R^2 = 0.0541 \end{aligned}$$

- Suppose $H_0 : \beta_{enroll} = 0$.
- Let $H_A : \beta_{enroll} \neq 0$.
- T-statistic?

$$t(\widehat{\beta}_{enroll} | \beta_{enroll} = 0) = \frac{\widehat{\beta}_{enroll} - 0}{se(\widehat{\beta}_{enroll})} = \frac{-0.0002}{0.00022} = -0.91$$

- $t_{crit} = 1.960$.
- Since $|-0.91| < 1.960$, we **cannot reject the null** that enrollment has no effect.


−1.960
−0.91
0
1.960

The t-test

Non-zero hypothesis

- Crime and School size

$$\log(Crime) = \beta_0 + \beta_1 \log(Enroll) + u$$

- We estimate:

$$\begin{aligned} \widehat{\log(Crime)} &= \underset{(0.104)}{-6.63} + \underset{(0.11)}{1.27} \log(Enroll) \\ obs &= 197, R^2 = 0.585 \end{aligned}$$

- Suppose $H_0 : \beta_1 = 1$
- How do we interpret this?
- 1% increase in enrollment causes a 1% increase in crime.
- T-stat:

$$t(\widehat{\beta}_2 | \beta_2 = 1) = \frac{\widehat{\beta}_2 - 1}{se(\widehat{\beta}_2)} = \frac{1.27 - 1}{0.11} = 2.45$$

The t-test

Non-zero hypothesis

- $H_A : \beta_1 \neq 1$
- Three steps
  1. Calculate $n - k - 1 = 197 - 1 - 1 = 195$
  2. At the 5% level, t_{crit} is roughly 1.96
  3. $|2.45| > 1.96 \Rightarrow$ reject H_0 in favor of H_A.
- A 1% increase in enrollment yields a greater than 1% increase in the crime rate.

Confidence intervals

A more informative technique

- Confidence intervals give us the region of most likely β_j.
- Recall that we cannot reject the null of a two sided t-test if:

$$\left| \frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \right| < t_{crit}$$

- This is satisfied if:

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} < t_{crit} \quad \text{and} \quad -\left(\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \right) < t_{crit}$$

- Manipulating the first, we get:

$$\widehat{\beta}_j - t_{crit} \cdot se(\widehat{\beta}_j) < \beta_j$$

- The second:

$$\beta_j < \widehat{\beta}_j + t_{crit} \cdot se(\widehat{\beta}_j)$$

Confidence intervals

A more informative technique

- Combining the two results, we get a "confidence interval" for β_j

$$\underbrace{\widehat{\beta}_j - t_{crit} \cdot se(\widehat{\beta}_j)}_{lower} < \beta_j < \underbrace{\widehat{\beta}_j + t_{crit} \cdot se(\widehat{\beta}_j)}_{upper}$$

- Range **which cannot be rejected** using a two-sided t-test.
- If t_{crit} is at calculated at 95% confidence, this is called the 95% confidence interval.
- Why are these more informative?

Confidence intervals

A more informative technique

- Combining the two results, we get a "confidence interval" for β_j

$$\underbrace{\widehat{\beta}_j - t_{crit} \cdot se(\widehat{\beta}_j)}_{lower} < \beta_j < \underbrace{\widehat{\beta}_j + t_{crit} \cdot se(\widehat{\beta}_j)}_{upper}$$

- Effect of $se(\widehat{\beta}_j)$ on confidence interval?
  - $se(\widehat{\beta}_j) \uparrow \Longrightarrow$ wider confidence interval.
- Effect of t_{crit}?
  - $t_{crit} \uparrow$ implies a higher level of confidence
  - $\Longrightarrow$ wider confidence interval
  - What if I desire 100% confidence?
    - $-\infty < \beta_j < \infty$

Confidence intervals

A more informative technique

- Crime example:

$$\begin{aligned} \widehat{\log(Crime)} &= \underset{(0.104)}{-6.63} + \underset{(0.11)}{1.27}\log(Enroll) \\ obs &= 197,\ R^2 = 0.585 \end{aligned}$$

- 95% confidence interval for the coefficient on $\log(Enroll)$?
- Interpret?

$$\begin{aligned} 1.27 - 1.96 \cdot 0.11 &< \beta_{enroll} < 1.27 + 1.96 \cdot 0.11 \\ 1.054 &< \beta_{enroll} < 1.485 \end{aligned}$$

- 99% Confidence interval?
- $T_{stat} = 2.58$
- Use Formula:

$$\begin{aligned} 1.27 - 2.58 \cdot 0.11 &< \beta_{enroll} < 1.27 + 2.58 \cdot 0.11 \\ 0.9862 &< \beta_{enroll} < 1.5538 \end{aligned}$$

Confidence intervals

Housing prices and cancer risk

- Housing prices and cancer risk (Davis, 2004)
- Housing Data - two similar Nevada Counties
  1. Churchill county
     - 31 new cases of Pediatric Leukemia (PL), 1997-2002
     - Similarly sized counties should expect 1 new case.
  2. Lyon county
     - A similar county, located directly to the west.
- Specification

$$\log(Price) = \beta_1 Risk + Other + u$$

  1. $\log(Price)$: log value of housing prices
  2. $Risk$: Index of perceived PL risk
  3. $Other$: Other controls (lot size, floor space, etc.)

- Housing data from 1990-2002

Confidence intervals

Housing prices and cancer risk

- Estimation (other effects suppressed for brevity)

$$\begin{aligned} \log(Price) &= \underset{(0.017)}{-0.156 Risk} \\ n &= 10204,\ R^2 = 0.63 \end{aligned}$$

- What is the 99% confidence interval β_1?
- $t_{crit} = 2.576$
- Use the formula:

$$\begin{aligned} \widehat{\beta}_1 - t_{crit} \cdot se(\widehat{\beta}_1) &< \beta_1 < \widehat{\beta}_1 + t_{crit} \cdot se(\widehat{\beta}_1) \\ -0.156 - 2.576 \cdot 0.017 &< \beta_1 < -0.156 + 2.576 \cdot 0.017 \\ -0.200 &< \beta_1 < -0.112 \end{aligned}$$

- At the 99% level, cancer risk negatively affects housing prices.

P-values

- How likely is it that I falsely reject the null?
- P-value:

  The probability that the null hypothesis is falsely rejected

- In the crime and enrollment example

$$H_0 : \beta_{Enroll} = 1,\ H_A : \beta_{Enroll} \neq 1$$

- T-statistic

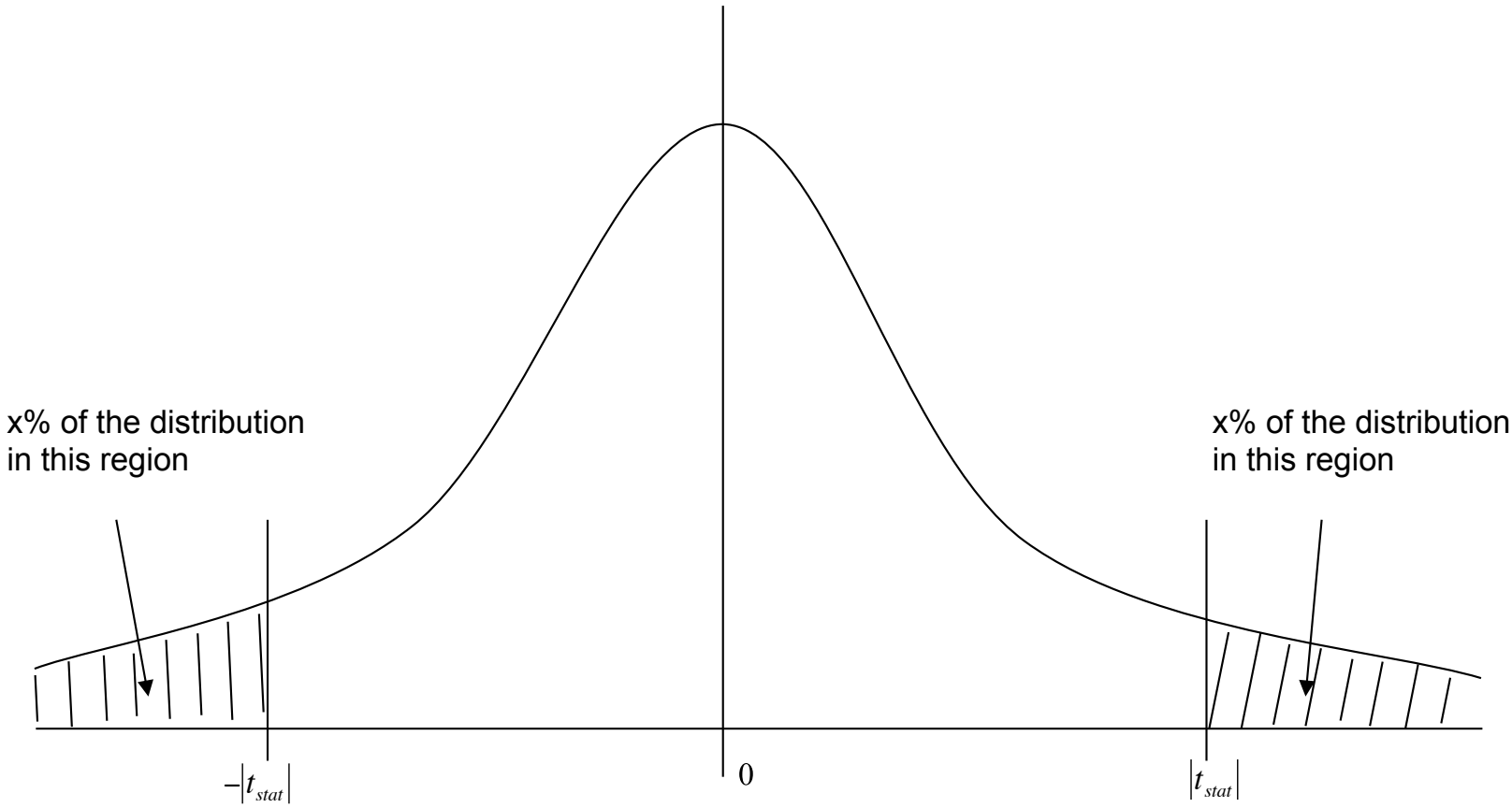
$$tstat = \frac{1.27 - 1}{0.11} = 2.45$$

- The p-value is the value of the following

$$pvalue = \Pr(|T| > 2.45)$$

- The p-value is the probability that I randomly draw a value from the t-distribution that is larger (in absolute terms) than the estimated T-statistic

P-value is the area of the shaded regions put together



P-values

- To find the p-value
  1. Calculate the t-statistic
  2. Find the closest match to your t-statistic on the normal table
  3. The p-value is the significance level that generates this t-statistic.
- In the crime and enrollment example
  1. $t_{stat} = 2.45$
  2. $\Pr(|T| > 2.45)$?

$$\begin{aligned}
\Pr(|T| &> 2.45) = \Pr(T > 2.45 \bigcup T < -2.45) \\
&= \Pr(T > 2.45 \bigcup T < -2.45) \\
&= \Pr(T > 2.45) + \Pr(T < -2.45) = 2 * \Pr(T > 2.45) \\
&= 2 * (1 - 0.9929) = 0.0142
\end{aligned}$$

- Thus, we will falsely reject the null less than 1.42% of the time.