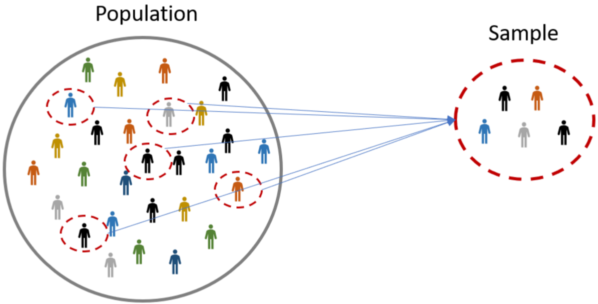
BS2280 - Econometrics 1

Lecture 4 - Part 1: Review of Estimation and Hypothesis Testing

Dr. Yichen Zhu

Structure of today's lecture

Background



	Population Parameters	**Inference**	**Sample Coefficients**
Statistics:	μ, σ^2	$\longleftarrow$	$\bar{x}, s^2$
Econometrics:	β_1, β_2	$\longleftarrow$	$\hat{\beta}_1, \hat{\beta}_2$

Background

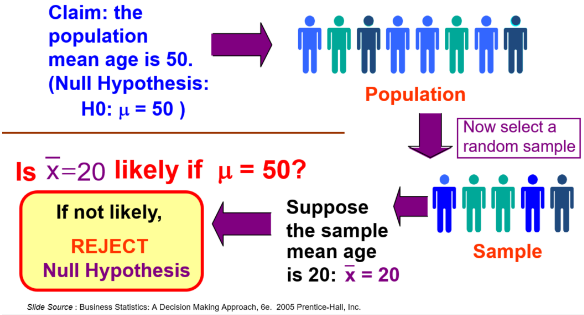
- What are point and interval estimates?
- What is hypothesis testing?
- Make sure to revise these topics, as they will be fundamental for regression analysis
- Review the hypothesis testing in statistics first

What is a Hypothesis?

- A hypothesis is a claim (assumption) about a population parameter (e.g. population mean μ)
- For example:
  - Claim: The mean age of people on the University Campus is $\mu = 50$

Null Hypothesis	**Alternative Hypothesis**
Begin with the assumption that the null hypothesis is true Example The mean age of people on the University Campus is 50 $H_0 : \mu = 50$	**The opposite of the null hypothesis** The mean age of people on the University Campus is not 50 H_A or H_1: $\mu \neq 50$

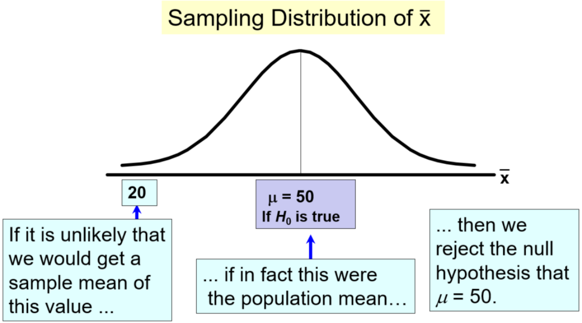
Hypothesis Testing Process



- If the sample mean $\bar{x} = 20$, how likely is population mean $\mu = 50$, say how likely is our claim correct?
- Based on the sample evidence of $\bar{x} = 20$, the null hypothesis of the population mean $\mu = 50$ is not realistic, we will reject the null hypothesis of the population mean $\mu = 50$.

Reason for Rejection H_0

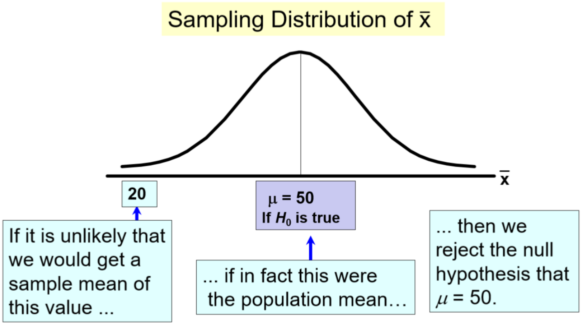
Question: How do we decide to reject the null hypothesis of population mean $\mu = 50$.



- How far is sample mean 20 away from the population mean 50?
- If sample mean 20 is far from population mean 50, reject; If sample mean 20 is not far from population mean 50, cannot reject.
- Hypothesis testing will find out when are we close enough to the population mean, then cannot reject.
- Now we need a threshold to tell us.

Reason for Rejection H_0

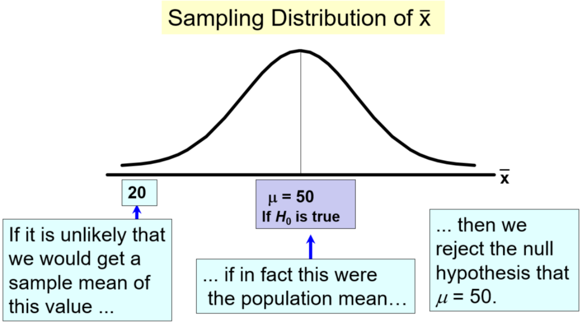
Question: How do we decide to reject the null hypothesis of population mean $\mu = 50$.



- How far is sample mean 20 away from the population mean 50?
- If sample mean 20 is far from population mean 50, reject; If sample mean 20 is not far from population mean 50, cannot reject.
- Hypothesis testing will find out when are we close enough to the population mean, then cannot reject.
- Now we need a threshold to tell us.

Reason for Rejection H_0

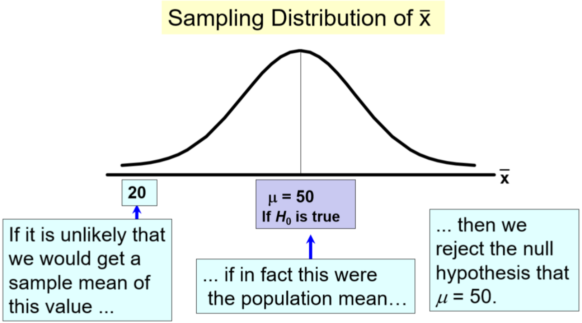
Question: How do we decide to reject the null hypothesis of population mean $\mu = 50$.



- How far is sample mean 20 away from the population mean 50?
- If sample mean 20 is far from population mean 50, reject; If sample mean 20 is not far from population mean 50, cannot reject.
- Hypothesis testing will find out when are we close enough to the population mean, then cannot reject.
- Now we need a threshold to tell us.

Reason for Rejection H_0

Question: How do we decide to reject the null hypothesis of population mean $\mu = 50$.

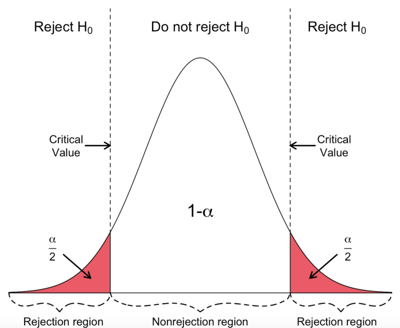


- How far is sample mean 20 away from the population mean 50?
- If sample mean 20 is far from population mean 50, reject; If sample mean 20 is not far from population mean 50, cannot reject.
- Hypothesis testing will find out when are we close enough to the population mean, then cannot reject.
- Now we need a threshold to tell us.

Level of Significance and Critical Value

Threshold: level of significance α and critical value(s).

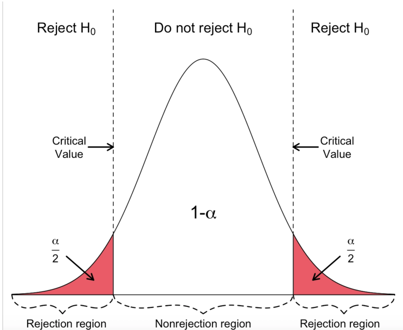
- Level of significance α: typical values are 1%, 5%, or 10%
- Defines unlikely values of sample statistic if null hypothesis is true
- Critical value(s) for a specified level of significance α is from a table or computer
- Defines rejection region of the sampling distribution



Level of Significance and Critical Value

Threshold: level of significance α and critical value(s).

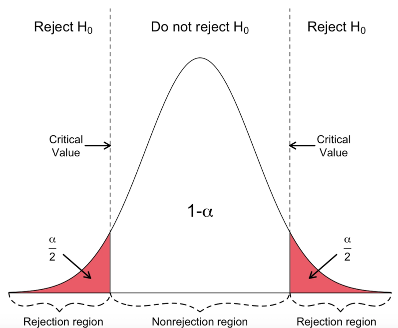
- Level of significance α: typical values are 1%, 5%, or 10%
- Defines unlikely values of sample statistic if null hypothesis is true
- Critical value(s) for a specified level of significance α is from a table or computer
- Defines rejection region of the sampling distribution



Level of Significance and Critical Value

Threshold: level of significance α and critical value(s).

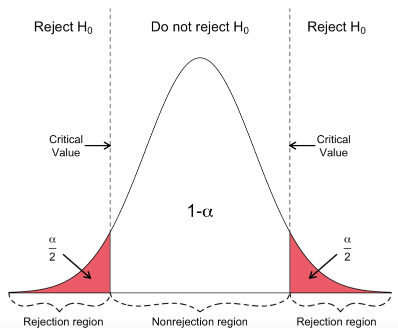
- Level of significance α: typical values are 1%, 5%, or 10%
- Defines unlikely values of sample statistic if null hypothesis is true
- Critical value(s) for a specified level of significance α is from a table or computer
- Defines rejection region of the sampling distribution



Level of Significance and Critical Value

Threshold: level of significance α and critical value(s).

- Level of significance α: typical values are 1%, 5%, or 10%
- Defines unlikely values of sample statistic if null hypothesis is true
- Critical value(s) for a specified level of significance α is from a table or computer
- Defines rejection region of the sampling distribution

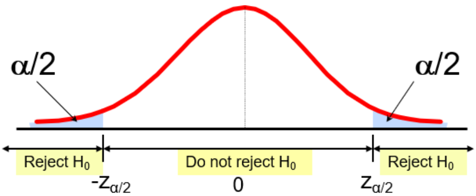


Critical Value Approach to Testing

- Cannot use sample statistics because different samples can be very different
- In our sample, sample mean $\bar{x} = 20$ represents years
  For the marks of econometrics exam, sample mean is about exam marks

Solution

- Convert sample statistic (e.g.: $\bar{x}$) to test statistic (z or t statistic)
- Do not use sample statistic (e.g.: $\bar{x}$), but use test statistic (z or t statistic)
- If the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0
- There are two cutoff values (critical values): $\pm z_{\frac{\alpha}{2}}$

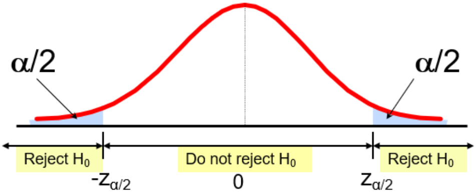


Critical Value Approach to Testing

- Cannot use sample statistics because different samples can be very different
- In our sample, sample mean $\bar{x} = 20$ represents years
  For the marks of econometrics exam, sample mean is about exam marks

Solution

- Convert sample statistic (e.g.: $\bar{x}$) to test statistic (z or t statistic)
- Do not use sample statistic (e.g.: $\bar{x}$), but use test statistic (z or t statistic)
- If the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0
- There are two cutoff values (critical values): $\pm z_{\frac{\alpha}{2}}$

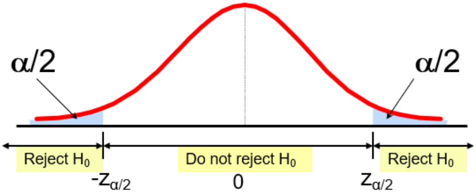


Critical Value Approach to Testing

- Cannot use sample statistics because different samples can be very different
- In our sample, sample mean $\bar{x} = 20$ represents years
  For the marks of econometrics exam, sample mean is about exam marks

Solution

- Convert sample statistic (e.g.: $\bar{x}$) to test statistic (z or t statistic)
- Do not use sample statistic (e.g.: $\bar{x}$), but use test statistic (z or t statistic)
- If the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0
- There are two cutoff values (critical values): $\pm z_{\frac{\alpha}{2}}$

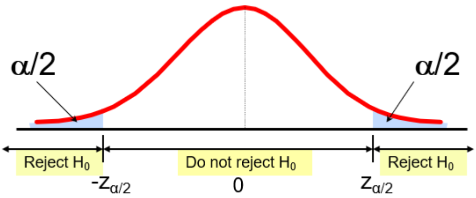


Critical Value Approach to Testing

- Cannot use sample statistics because different samples can be very different
- In our sample, sample mean $\bar{x} = 20$ represents years
  For the marks of econometrics exam, sample mean is about exam marks

Solution

- Convert sample statistic (e.g.: $\bar{x}$) to test statistic (z or t statistic)
- Do not use sample statistic (e.g.: $\bar{x}$), but use test statistic (z or t statistic)
- If the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0
- There are two cutoff values (critical values): $\pm z_{\frac{\alpha}{2}}$

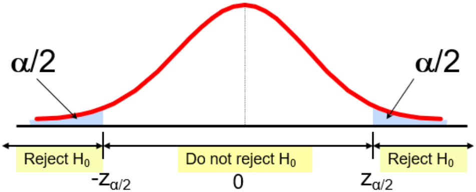


Critical Value Approach to Testing

- Cannot use sample statistics because different samples can be very different
- In our sample, sample mean $\bar{x} = 20$ represents years
  For the marks of econometrics exam, sample mean is about exam marks

Solution

- Convert sample statistic (e.g.: $\bar{x}$) to test statistic (z or t statistic)
- Do not use sample statistic (e.g.: $\bar{x}$), but use test statistic (z or t statistic)
- If the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0
- There are two cutoff values (critical values): $\pm z_{\frac{\alpha}{2}}$



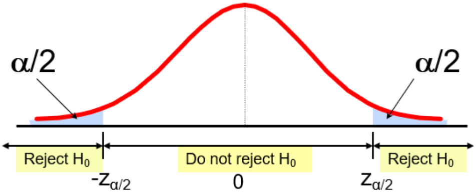
Critical Value Approach to Testing

- Cannot use sample statistics because different samples can be very different
- In our sample, sample mean $\bar{x} = 20$ represents years

  For the marks of econometrics exam, sample mean is about exam marks

Solution

- Convert sample statistic (e.g.: $\bar{x}$) to test statistic (z or t statistic)
- Do not use sample statistic (e.g.: $\bar{x}$), but use test statistic (z or t statistic)
- If the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0
- There are two cutoff values (critical values): $\pm z_{\frac{\alpha}{2}}$



Hypothesis testing when population variance is unknown

- Usually, the population variance is unknown. Therefore, we will use the t-distribution for hypothesis tests
- t-statistic: it measures the number of standard errors some sample mean is from an expected mean

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}$$

- If the t-statistic is further away from the mean than the critical value, we can reject the null hypothesis

Example: Two-Tail Test (σ unknown)

The average cost of a hotel room in New York is said to be \$168 per night. A random sample of 25 hotels resulted in $\bar{x} = \$172.50$ and $s = \$15.40$. Test at the $\alpha = 0.05$ level. Assume the population distribution is normal.

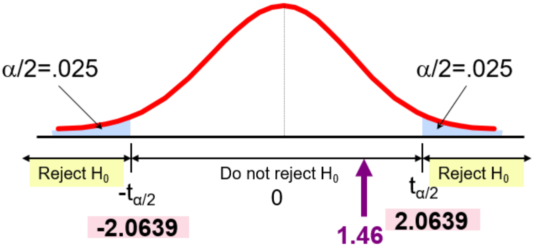
- $H_0 : \mu = 168$
- $H_1 : \mu \neq 168$

Steps for hypothesis testing:

1. State the null and alternative hypotheses
2. Select the significance level
3. Select and calculate the test statistics
4. Set the decision rule
5. Make statistical decisions

Example Solution: Two-Tail Test

H_0: $\mu = 168$
H_A: $\mu \neq 168$



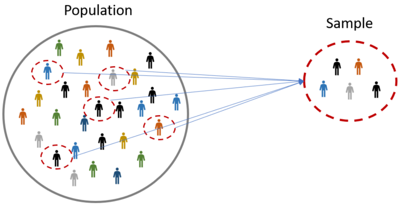
- $\alpha = 0.05$
- n = 25
- σ is unknown, so use a **t statistic**
- **Critical Value:**

$t_{24} = \pm 2.0639$

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

Do not reject H$_0$: not sufficient evidence that true mean cost is different than $168

Estimation Background



What is Estimation?

Estimation is the process of using sample data to draw inferences about the population. For example, we use OLS estimator to estimate β_1 and β_2

	Population Parameters	**Inference**	**Sample Coefficients**
Statistics:	μ, σ^2	$\longleftarrow$	$\bar{x}, s^2$
Econometrics:	β_1, β_2	$\longleftarrow$	$\hat{\beta}_1, \hat{\beta}_2$

What about hypothesis testing for regressions?

	Statistics	Econometrics
Model	X, unknown μ, σ^2	$Y = \beta_1 + \beta_2 X + u$
Estimator	$\bar{X}$ and s^2	$\hat{\beta}_1$ and $\hat{\beta}_2$
Null Hypothesis **Alternative Hypothesis**	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \beta_1 = \beta_1^0$ $H_0 : \beta_2 = \beta_2^0$ $H_1 : \beta_1 \neq \beta_1^0$ $H_1 : \beta_2 \neq \beta_2^0$
Test statistic	$t = \frac{\bar{X} - \mu_0}{s.e.(\bar{X})}$	$t = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)}$ $t = \frac{\hat{\beta}_2 - \beta_2^0}{s.e.(\hat{\beta}_2)}$
Reject H_0 if	$\lvert t \rvert > t_{crit}$	$\lvert t \rvert > t_{crit}$
Degrees of Freedom	$n - 1$	$n - k = n - 2$

Point Estimate

- A point estimator of a population parameter is a function of the sample information that yields a single number.
- The corresponding realisation is called the point estimate of the parameter.
- The point estimate is a single value.
- For Example:
- The coefficient $\hat{\beta}_2$ derived through OLS is assumed to be an unbiased and most efficient estimate for parameter β_2, if our OLS assumptions hold!

Interval Estimate

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- **Interval estimate**: a range of values, expressing the degree of uncertainty.
- For Example: We are 95% confident that true effect of years of education on hourly wages is between USD 0.9 and USD 1.6.
- Such interval estimates are called confidence intervals.
- Confidence intervals are usually set at the 95% level.
- **This can be interpreted as follows:** 95% chance that the population parameter will lie in the calculated intervals.

Interval Estimate

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- **Interval estimate**: a range of values, expressing the degree of uncertainty.
- For Example: We are 95% confident that true effect of years of education on hourly wages is between USD 0.9 and USD 1.6.
- Such interval estimates are called confidence intervals.
- Confidence intervals are usually set at the 95% level.
- **This can be interpreted as follows:** 95% chance that the population parameter will lie in the calculated intervals.

Interval Estimate

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- **Interval estimate**: a range of values, expressing the degree of uncertainty.
- For Example: We are 95% confident that true effect of years of education on hourly wages is between USD 0.9 and USD 1.6.
- Such interval estimates are called confidence intervals.
- Confidence intervals are usually set at the 95% level.
- **This can be interpreted as follows:** 95% chance that the population parameter will lie in the calculated intervals.

Interval Estimate

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- **Interval estimate**: a range of values, expressing the degree of uncertainty.
- For Example: We are 95% confident that true effect of years of education on hourly wages is between USD 0.9 and USD 1.6.
- Such interval estimates are called confidence intervals.
- Confidence intervals are usually set at the 95% level.
- **This can be interpreted as follows:** 95% chance that the population parameter will lie in the calculated intervals.

Interval Estimate

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- **Interval estimate**: a range of values, expressing the degree of uncertainty.
- For Example: We are 95% confident that true effect of years of education on hourly wages is between USD 0.9 and USD 1.6.
- Such interval estimates are called confidence intervals.
- Confidence intervals are usually set at the 95% level.
- **This can be interpreted as follows:** 95% chance that the population parameter will lie in the calculated intervals.

Interval Estimate

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- **Interval estimate**: a range of values, expressing the degree of uncertainty.
- For Example: We are 95% confident that true effect of years of education on hourly wages is between USD 0.9 and USD 1.6.
- Such interval estimates are called confidence intervals.
- Confidence intervals are usually set at the 95% level.
- **This can be interpreted as follows:** 95% chance that the population parameter will lie in the calculated intervals.

Interval Estimate

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- **Interval estimate**: a range of values, expressing the degree of uncertainty.
- For Example: We are 95% confident that true effect of years of education on hourly wages is between USD 0.9 and USD 1.6.
- Such interval estimates are called confidence intervals.
- Confidence intervals are usually set at the 95% level.
- **This can be interpreted as follows:** 95% chance that the population parameter will lie in the calculated intervals.

Confidence Intervals

Model $Y = \beta_1 + \beta_2 X + u$

Null hypothesis $H_0 : \beta_2 = \beta_2^0$

Alternative hypothesis $H_1 : \beta_2 \neq \beta_2^0$

Reject H_0 if $\dfrac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}(\hat{\beta}_2)} > t_{\text{crit}}$ **or** $\dfrac{\hat{\beta}_2 - \beta_2^0}{\text{s.e.}(\hat{\beta}_2)} < -t_{\text{crit}}$

Reject H_0 if $\hat{\beta}_2 - \beta_2^0 > \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$ **or** $\hat{\beta}_2 - \beta_2^0 < -\text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$

Reject H_0 if $\hat{\beta}_2 - \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} > \beta_2^0$ **or** $\hat{\beta}_2 + \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} < \beta_2^0$

Do not reject H_0 if $\hat{\beta}_2 - \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}} \leq \beta_2 \leq \hat{\beta}_2 + \text{s.e.}(\hat{\beta}_2) \times t_{\text{crit}}$

Confidence Intervals using R

```
> summary(earnfit)

Call:
lm(formula = EARNINGS ~ S, data = EAWE21.simple)

Residuals:
Min      1Q  Median      3Q     Max
-20.079  -6.726  -2.203   3.451  79.037

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.7647     2.8038   0.273    0.785
S              1.2657     0.1855   6.824 2.58e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.36 on 498 degrees of freedom
Multiple R-squared:  0.08551,Adjusted R-squared:  0.08368
F-statistic: 46.57 on 1 and 498 DF,  p-value: 2.579e-11
```

Confidence Intervals using R

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.7647     2.8038   0.273    0.785
S              1.2657     0.1855   6.824 2.58e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\hat{\beta}_2 - s.e.(\hat{\beta}_2) \times t_{crit} \leq \beta_2 \leq \hat{\beta}_2 + s.e.(\hat{\beta}_2) \times t_{crit}$$

- The point estimate $\hat{\beta}_2$ is 1.266 and its standard error is 0.185:

$$1.266 - 0.185 \times 1.965 \leq \beta_2 \leq 1.266 + 0.185 \times 1.965$$
$$0.902 \leq \beta_2 \leq 1.630$$

- 95% confidence interval means: 95% chance that the population parameter β_2 will lie in [0.902,1.630].

Confidence Intervals using R

```
> confint(earnfit)
                  2.5 %   97.5 %
(Intercept) -4.7439819 6.273351
S            0.9012959 1.630128
```

- `confint` command uses regression output as argument
- Results confirm manual calculations!