UNIVERSITÀ DEGLI STUDI DELLA TUSCIA

SAPIENZA UNIVERSITÀ DI ROMA

# Statistics for business and decision making

Dr. Ilaria Benedetti

## 10. Simple regression model - Interval estimation and Hypothesis tests

# Interval Estimation

There are two types of estimates

- **Point estimates**
  - The estimate $b_2$ is a point estimate of the unknown population parameter in the regression model.
- **Interval estimates**
  - Interval estimation proposes a range of values in which the true parameter is likely to fall
  - Providing a range of values gives a sense of what the parameter value might be, and the precision with which we have estimated it
  - Such intervals are often called **confidence intervals**.
    - We prefer to call them **interval estimates** because the term ‘‘confidence’’ is widely misunderstood and misused

- The normal distribution of $b_2$, the least squares estimator of $\beta_2$, is

$$b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

- A standardized normal random variable is obtained from $b_2$ by subtracting its mean and dividing by its standard deviation:

$$Z = \frac{b2 - \beta_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \sim N(0,1)$$

- We know that:

$$P(-1.96 \le Z \le 1.96) = 0.95$$

- Substituting:

$$P\left(-1.96 \le \frac{b2 - \beta_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \le 1.96\right) = 0.95$$

# Rearranging:

$$P\left(b_2 - 1.96\sqrt{\sigma^2/\sum(x_i-\bar{x})^2} \le \beta_2 \le b_2 + 1.96\sqrt{\sigma^2/\sum(x_i-\bar{x})^2}\right) = 0.95$$

- The two end-points $b_2 \pm 1.96\sqrt{\sigma^2/\sum(x_i-\bar{x})^2}$ provide an ***interval estimator***.

- In repeated sampling 95% of the intervals constructed this way will contain the true value of the parameter $\beta_2$.
- This easy derivation of an interval estimator is based on the assumption SR6 *and* that we know the variance of the error term $\sigma^2$.

Replacing $\sigma^2$ with $\hat{\sigma}^2$ creates a **random variable *t:***

$$t = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum (x_i - \bar{x})^2}} = \frac{b_2 - \beta_2}{\sqrt{\hat{\text{var}}(b_2)}} = \frac{b_2 - \beta_2}{se(b_2)} \sim t_{(N-2)}$$

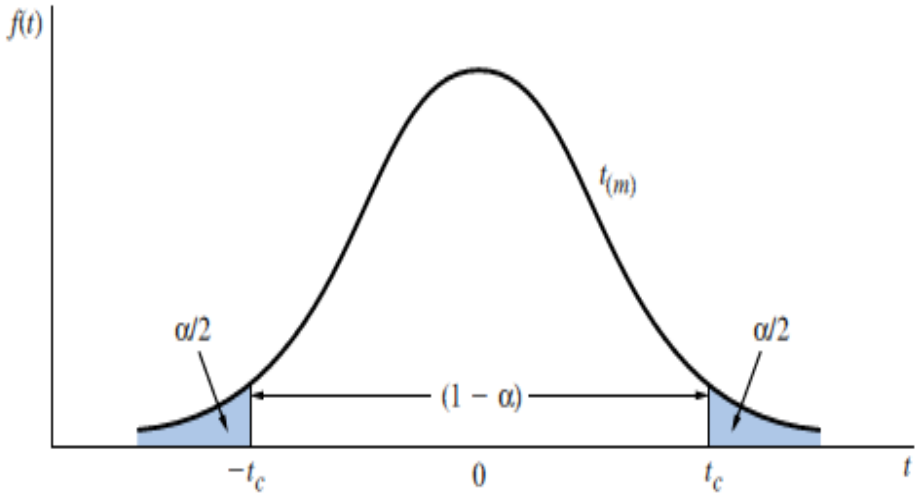
The ratio $\boldsymbol{t = \frac{b_2 - \beta_2}{se(b_2)}}$ has a *t*-distribution with $(N-2)$ degrees of freedom, which we denote as: $t \sim t(n-2)$

In general we can say, if assumptions SR1-SR6 hold in the simple linear regression model, then

$$\boldsymbol{t = \frac{b_2 - \beta_2}{se(b_2)} \sim t(n-2) \, for \; k = 1, 2}$$

# *Student t- distribution*

- The *t*-distribution is a bell shaped curve centered at zero;
- It looks like the standard normal distribution, except it is ***more spread out***, with a larger variance and thicker tails;
- The shape of the *t*-distribution is controlled by a single parameter called the **degrees of freedom**, often abbreviated as *df*



We can find a "**critical value**" from a t-distribution such that

$$P(t \geq t_c) = P(t \leq -t_c) = \alpha/2$$

where **α** is a probability often taken to be α = 0.01 or α = 0.05. The critical value $t_c$ for degrees of freedom *m* is the percentile value $t_{(1-\alpha/2,\, m)}$

Each shaded ''tail'' area contains α/2 of the probability, so that 1-α of the probability is contained in the center portion. Consequently, we can make the probability statement

$$P(-t_c \leq t \leq t_c) = 1 - \alpha$$

or

$$P\left(-t_c \leq \frac{b_k - \beta_k}{se(b_k)} \leq t_c\right) = 1 - \alpha$$

or

$$P[b_k - t_c se(b_k) \leq \beta_k \leq t_c + t_c se(b_k)] = 1 - \alpha$$

- When $b_k$ and se($b_k$) are estimated values (numbers), based on a given sample of data, then $b_k \pm t_c se(b_k)$ is called a 100(1-α)% interval estimate of $b_k$.
  - Equivalently it is called a 100(1-α)% confidence interval.
  - Usually α = 0.01 or α = 0.05, so that we obtain a 99% confidence interval or a 95% confidence interval.

- ***The interpretation of confidence intervals requires a great deal of care***
  - The properties of the interval estimation procedure are based on the notion of **repeated sampling**
  - Any one interval estimate, based on one sample of data, may or may not contain the true parameter $\beta_k$, and because $\beta_k$ is unknown, we will never know whether it does or does not
  - When ‘‘confidence intervals’’ are discussed, remember that our confidence is in the procedure used to construct the interval estimate; it is not in any one interval estimate calculated from a sample of data

```
> summary(lm)

Call:
lm(formula = food_exp ~ income, data = food)

Residuals:
     Min       1Q   Median       3Q      Max
-223.025  -50.816   -6.324   67.879  212.044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.416     43.410   1.922   0.0622 .
income        10.210      2.093   4.877 1.95e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.52 on 38 degrees of freedom
Multiple R-squared:  0.385,     Adjusted R-squared:  0.3688
F-statistic: 23.79 on 1 and 38 DF,  p-value: 1.946e-05
```

$se(b_2)$

$$t = \frac{b_2}{se(b_2)}$$

# Example on food expenditure data

$$P\left[b_2 - 2.024se(b_2) \le \beta_2 \le b_2 + 2.024se(b_2)\right] = 0.95$$

The critical value $t_c = 2.024$, which is appropriate for $\alpha = .05$ and 38 degrees of freedom

To construct an interval estimate for $\beta_2$ we use the least squares estimate $b_2 = 10.21$ and its standard error

$$se(b_2) = \sqrt{\hat{\text{var}}(b_2)} = \sqrt{4.38} = 2.09$$

A **"95% confidence interval estimate" for $\beta_2$**:

$$b_2 \pm t_c se(b_2) = 10.21 \pm 2.024(2.09) = [5.97, 14.45]$$

When the procedure we used is applied to many random samples of data from the same population, then 95% of all the interval estimates constructed using this procedure will contain the true parameter

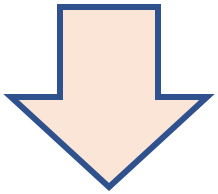
# Is $\beta_2$ actually in the interval [5.97, 14.45]?

- We do not know, and we will never know…
- What we do know is that when the procedure we used is applied to many random samples of data from the same population, then 95% of all the interval estimates constructed using this procedure will contain the true parameter
- The interval estimation procedure ‘‘works’’ 95% of the time
- What we can say about the interval estimate based on our one sample is that, given the reliability of the procedure, we would be ‘‘surprised’’ if $\beta_2$ is not in the interval [*5.97, 14.45*].

# *What is the usefulness of an interval estimate of $\beta_2$?*

When reporting regression results, we always give a point estimate, such as $b_2 = 10.21$

However, the point estimate alone gives no sense of its reliability



Thus, we might also report an interval estimate

***Interval estimates incorporate both the point estimate and the standard error of the estimate, which is a measure of the variability of the least squares estimator***

# Example on food expenditure data

| Sample | $b_1$ | se($b_1$) | $b_2$ | se($b_2$) | $\hat{\sigma}^2$ |
|---|---|---|---|---|---|
| 1 | 131.69 | 40.58 | 6.48 | 1.96 | 7002.85 |
| 2 | 57.25 | 33.13 | 10.88 | 1.60 | 4668.63 |
| 3 | 103.91 | 37.22 | 8.14 | 1.79 | 5891.75 |
| 4 | 46.50 | 33.33 | 11.90 | 1.61 | 4722.58 |
| 5 | 84.23 | 41.15 | 9.29 | 1.98 | 7200.16 |
| 6 | 26.63 | 45.78 | 13.55 | 2.21 | 8911.43 |
| 7 | 64.21 | 32.03 | 10.93 | 1.54 | 4362.12 |
| 8 | 79.66 | 29.87 | 9.76 | 1.44 | 3793.83 |
| 9 | 97.30 | 29.14 | 8.05 | 1.41 | 3610.20 |
| 10 | 95.96 | 37.18 | 7.77 | 1.79 | 5878.71 |

# Example on food expenditure data

*Confidence interval computation*

| Sample | $b_1 - t_c \text{se}(b_1)$ | $b_1 + t_c \text{se}(b_1)$ | $b_2 - t_c \text{se}(b_2)$ | $b_2 + t_c \text{se}(b_2)$ |
|---|---|---|---|---|
| 1 | 49.54 | 213.85 | 2.52 | 10.44 |
| 2 | −9.83 | 124.32 | 7.65 | 14.12 |
| 3 | 28.56 | 179.26 | 4.51 | 11.77 |
| 4 | −20.96 | 113.97 | 8.65 | 15.15 |
| 5 | 0.93 | 167.53 | 5.27 | 13.30 |
| 6 | −66.04 | 119.30 | 9.08 | 18.02 |
| 7 | −0.63 | 129.05 | 7.81 | 14.06 |
| 8 | 19.19 | 140.13 | 6.85 | 12.68 |
| 9 | 38.32 | 156.29 | 5.21 | 10.89 |
| 10 | 20.69 | 171.23 | 4.14 | 11.40 |

# Hypothesis Tests

Hypothesis testing procedures compare a conjecture we have about a population to the information contained in a sample of data

- Given an economic and statistical model, hypotheses are formed about economic behavior.
- These hypotheses are then represented as statements about model parameters
- Hypothesis tests use the information about a parameter that is contained in a sample of data, its least squares point estimate, and its standard error, to draw a conclusion about the hypothesis

# Hypothesis Tests

## *Steps to follow*

1. A null hypothesis $H_0$
2. An alternative hypothesis $H_1$
3. A test statistic
4. A rejection region
5. A conclusion

# *Null vs alternative hypothesis*

A null hypothesis is the belief we will maintain until we are convinced by the sample evidence that it is not true, in which case we reject the null hypothesis

The null hypothesis is stated as $H_0: \beta_k = c$, where $c$ is a constant, and is an important value in the context of a specific regression model. A common value for $c$ is 0

Paired with every null hypothesis is a logical alternative hypothesis H1 that we will accept if the null hypothesis is rejected

The alternative hypothesis is flexible and depends to some extent on economic theory.

$$H_1: \beta_k > c$$
$$H_1: \beta_k < c$$
$$H_1: \beta_k \neq c$$

Based on the value of a test statistic we decide either to reject the null hypothesis or not to reject it

- **A test statistic has a special characteristic**: *its probability distribution is completely known when the null hypothesis is true, and it has some other distribution if the null hypothesis is not true*

The primary test statistic is:

$$t = \frac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(N-2)}$$

**If the null hypothesis H0 :bk= c is true**, then we can substitute c for bk and it follows that:

$$t = \frac{b_k - c}{\text{se}(b_k)} \sim t_{(N-2)}$$

**If the null hypothesis is not true**, then the t-statistic does not have a t-distribution with N-2 degrees of freedom.

# Rejection region

The rejection region depends on the form of the alternative.
It is the range of values of the test statistic that leads to rejection of the null hypothesis.

It is possible to construct a rejection region only if we have:

- A test statistic whose distribution is known when the null hypothesis is true
- An alternative hypothesis
- A level of significance

The rejection region consists of values that are unlikely and that have low probability of occurring when the null hypothesis is true. The chain of logic is:

> "*If a value of the test statistic is obtained that falls in a region of low probability, then it is unlikely that the test statistic has the assumed distribution, and thus it is unlikely that the null hypothesis is true*"

- If the alternative hypothesis is true, then values of the test statistic will tend to be unusually large or unusually small.

- The terms ‘‘large’’ and ‘‘small’’ are determined by choosing a probability α, called the **level of significance of the test**, which provides a meaning for ‘‘an unlikely event’’.

- The level of significance of the test α is usually chosen to be 0.01, 0.05 or 0.10

# *Types of error*

- If we reject the null hypothesis when it is true, then we commit what is called a ***Type I error***

  *The level of significance of a test is the probability of committing a Type I error*

  $P(\text{Type I error}) = \alpha$

- If we do not reject a null hypothesis that is false, then we have committed a ***Type II error***

***Do you reject the null hypothesis, or do you not reject the null hypothesis?***

- Avoid saying that you "***accept***" the null hypothesis.
- Make it standard practice to say what the conclusion means in the economic context of the problem you are working on and the economic significance of the finding

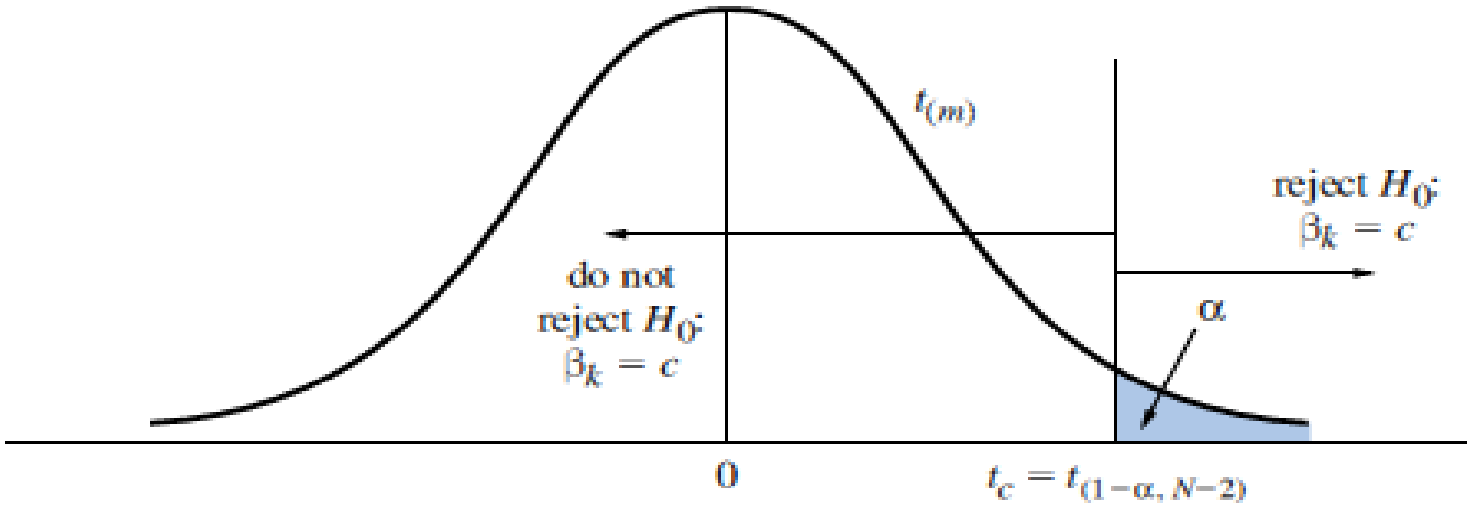
# Rejection Regions for Specific Alternatives

To have a rejection region for a null hypothesis, we need:

1. *A test statistic*
2. *A specific alternative*
3. *A level of significance, α, for the test*

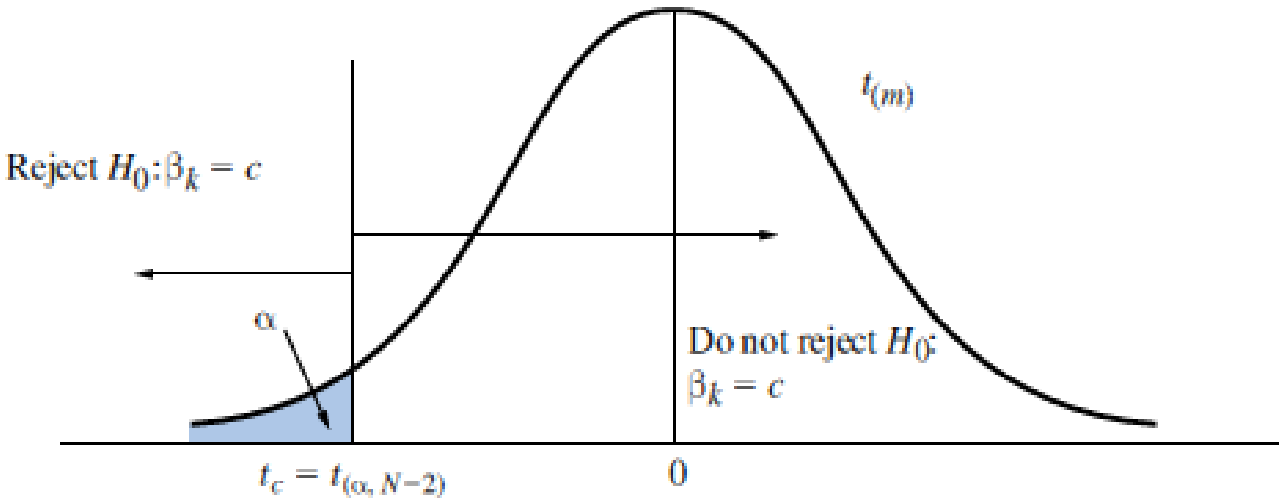
When testing the null hypothesis $H_0$:$\beta_k = c$ against the alternative hypothesis $H_1$:$\beta_k > c$, reject the null hypothesis and accept the alternative hypothesis if

$$t \geq t_{(1-\alpha;N-2)}$$



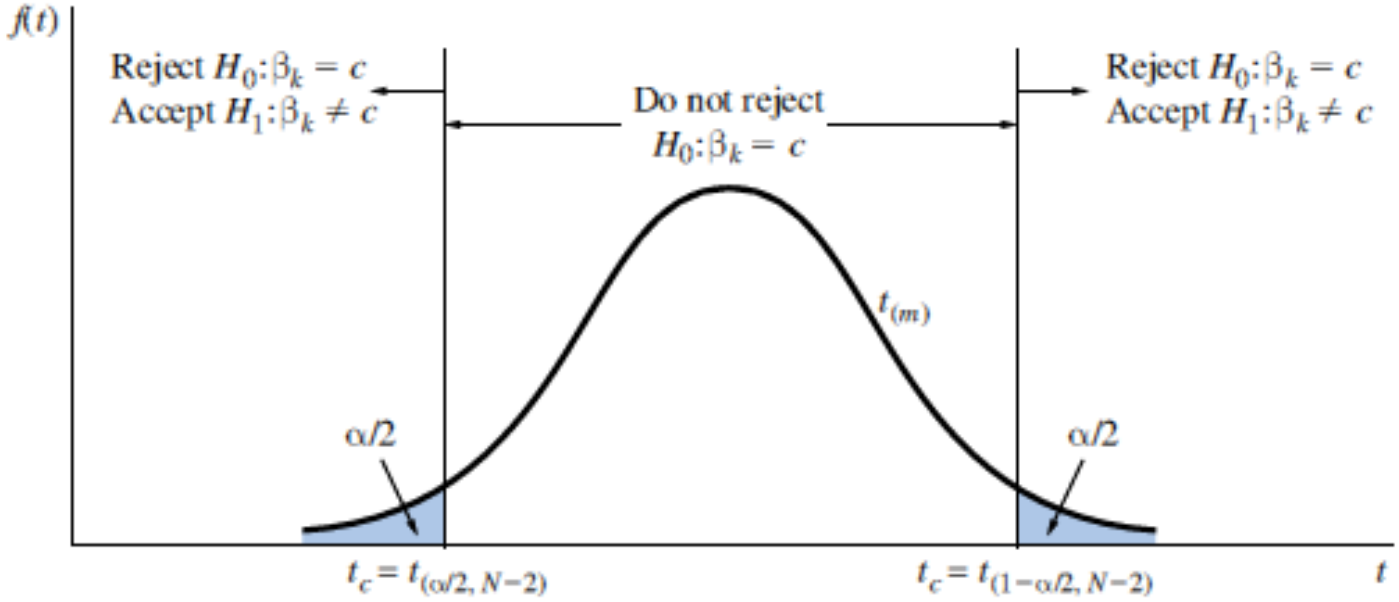
When testing the null hypothesis $H_0$:$\beta_k = c$ against the alternative hypothesis $H_1$:$\beta_k < c$, reject the null hypothesis and accept the alternative hypothesis if

$$t \leq t_{(1-\alpha;N-2)}$$



When testing the null hypothesis $H_0$:$\beta_k = c$ against the alternative hypothesis $H_1$:$\beta_k \neq c$, reject the null hypothesis and accept the alternative hypothesis if

$$t \leq t_{(1-\alpha;N-2)} \text{ or } t \geq t_{(1-\alpha;N-2)}$$



# *Examples of Hypothesis Tests*

1. Determine the null and alternative hypotheses.
2. Specify the test statistic and its distribution if the null hypothesis is true.
3. Select $\alpha$ and determine the rejection region.
4. Calculate the sample value of the test statistic.
5. State your conclusion.

# Examples of Hypothesis Tests

The null hypothesis is $H_0: \beta_2 = 0$

The alternative hypothesis is $H_1: \beta_2 > 0$

The test statistic is:

In this case $c = 0$, so $t = b_2/se(b_2) \sim t_{(N-2)}$ if the null hypothesis is true

Select $\alpha = 0.05$

The critical value for the right-tail rejection region is the 95th percentile of the $t$-distribution with $N - 2 = 38$ degrees of freedom, $t_{(0.95,38)} = 1.686$.

Thus we will reject the null hypothesis if the calculated value of $t \geq 1.686$.

If $t < 1.686$, we will not reject the null hypothesis

# Examples of Hypothesis Tests

Using the food expenditure data, we found that $b_2 = 10.21$ with standard error $\text{se}(b_2) = 2.09$

The value of the test statistic is: $$t = \frac{b_2}{\text{se}(b_2)} = \frac{10.21}{2.09} = 4.88$$

Since $t = 4.88 > 1.686$, we reject the null hypothesis that $\beta_2 = 0$ and accept the alternative that $\beta_2 > 0$

That is, ***we reject the hypothesis that there is no relationship between income and food expenditure***, and conclude that there is a *statistically significant* positive relationship between household income and food expenditure

# Examples of Hypothesis Tests

The null hypothesis is $H_0: \beta_2 = 5.5$

The alternative hypothesis is $H_1: \beta_2 > 5.5$

The test statistic is

$t = (b_2 - 5.5)/\text{se}(b_2) \sim t_{(N-2)}$ if the null hypothesis is true

Select $\alpha = 0.01$

The critical value for the right-tail rejection region is the 99th percentile of the $t$-distribution with $N - 2 = 38$ degrees of freedom, $t_{(0.99,38)} = 2.429$

Thus we will reject the null hypothesis if the calculated value of $t \geq 2.429$

If $t < 2.429$, we will not reject the null hypothesis

# Examples of Hypothesis Tests

Using the food expenditure data, we found that $b_2 = 10.21$ with standard error $\text{se}(b_2) = 2.09$

The value of the test statistic is:

$$t = \frac{b_2 - 5.5}{\text{se}(b_2)} = \frac{10.21 - 5.5}{2.09} = 2.25$$

Since $t = 2.25 < 2.429$ we do not reject the null hypothesis that $\beta_2 = 5.5$

We are *not* able to conclude that the new supermarket will be profitable and will not begin construction

# Examples of Hypothesis Tests

The null hypothesis is $H_0$:$\beta_2 = 15$
The alternative hypothesis is $H_1$:$\beta_2 < 15$

The test statistic is
$t = (b_2 - 15)/\text{se}(b_2) \sim t_{(N-2)}$ if the null hypothesis is true

Select $\alpha = 0.05$

The critical value for the left-tail rejection region is the 5th percentile of the $t$-distribution with $N - 2 = 38$ degrees of freedom, $t_{(0.05,38)} = -1.686$.

Thus we will reject the null hypothesis if the calculated value of $t \leq -1.686$
If $t > -1.686$, we will not reject the null hypothesis

# Examples of Hypothesis Tests

Using the food expenditure data, we found that $b_2 = 10.21$ with standard error $se(b_2) = 2.09$

The value of the test statistic is:

$$t = \frac{b_2 - 5.5}{se(b_2)} = \frac{10.21 - 15}{2.09} = -2.29$$

Since $t = -2.29 < -1.686$ we reject the null hypothesis that $\beta_2 = 15$ and accept the alternative that $\beta_2 < 15$

We conclude that households spend less than \$15 from each additional \$100 income on food

# *Examples of Hypothesis Tests*

The null hypothesis is $H_0: \beta_2 = 7.5$

The alternative hypothesis is $H_1: \beta_2 \neq 7.5$

The test statistic is

$t = (b_2 - 7.5)/\text{se}(b_2) \sim t_{(N-2)}$ if the null hypothesis is true

Select $\alpha = 0.05$

The critical value for the two-tail rejection region is the $2.5^{th}$ percentile of the *t*-distribution with $N - 2 = 38$ degrees of freedom, $t_{(0.025,38)} = -2.024$ and the $97.5^{th}$ percentile $t_{(0.975,38)} = 2.024$

Thus we will reject the null hypothesis if the calculated value of $t \geq 2.024$ **or** if $t \leq -2.024$

# Examples of Hypothesis Tests

Using the food expenditure data, we found that $b_2 = 10.21$ with standard error $se(b_2) = 2.09$

The value of the test statistic is

$$t = \frac{b_2 - 5.5}{se(b_2)} = \frac{10.21 - 7.5}{2.09} = 1.29$$

Since $-2.024 < t = 1.29 < 2.024$ we do not reject the null hypothesis that $\beta_2 = 7.5$

The sample data are consistent with the conjecture households will spend an additional \$7.50 per additional \$100 income on food.

# Examples of Hypothesis Tests

The null hypothesis is $H_0: \beta_2 = 0$

The alternative hypothesis is $H_1: \beta_2 \neq 0$

The test statistic is

$t = (b_2)/\text{se}(b_2) \sim t_{(N-2)}$ if the null hypothesis is true

Select $\alpha = 0.05$

The critical value for the two-tail rejection region is the $2.5^{th}$ percentile of the $t$-distribution with $N - 2 = 38$ degrees of freedom, $t_{(0.025,38)} = -2.024$ and the $97.5^{th}$ percentile $t_{(0.975,38)} = 2.024$

Thus we will reject the null hypothesis if the calculated value of $t \geq 2.024$ **or** if $t \leq -2.024$

# Examples of Hypothesis Tests

Using the food expenditure data, we found that $b_2 = 10.21$ with standard error se($b_2$) = 2.09

The value of the test statistic is

$$t = \frac{b_2}{\text{se}(b_2)} = \frac{10.21}{2.09} = 4.88$$

Since 4.88 > 2.024 we reject the null hypothesis that $\beta_2 = 0$
We conclude that there is a statistically significant relationship between income and food expenditure

From typical output, we can easily find the calculated *t* value used in this example

| Variable | Coefficient | Std. Error | *t*-Statistic | Prob. |
|---|---|---|---|---|
| *C* | 83.41600 | 43.41016 | 1.921578 | 0.0622 |
| *INCOME* | 10.20964 | 2.093264 | 4.877381 | 0.0000 |

# The p-Value

When reporting the outcome of statistical hypothesis tests, it has become standard practice to report the *p*-value (an abbreviation for probability value) of the test.

- If we have the ***p-value of a test, p***, we can determine the outcome of the test by comparing the *p*-value to the chosen level of significance, α, without looking up or calculating the critical values. This is much more convenient

Reject the null hypothesis when the p-value is less than,

or equal to, the level of significance α.

That is, if $p \leq \alpha$ then reject H0.

If $p > \alpha$ then do not reject H0.

# The p-Value

If $t$ is the calculated value of the $t$-statistic, then:

- $H_1: \beta_K > c$

  $p$ = probability to the right of $t$

- $H_1: \beta_K < c$

  $p$ = probability to the left of $t$

- $H_1: \beta_K \neq c$

  $p$ = sum of probabilities to the right of $|t|$ and to the left of $-|t|$

# The p-Value: examples

The null hypothesis is $H_0$: $\beta_2 = 5.5$

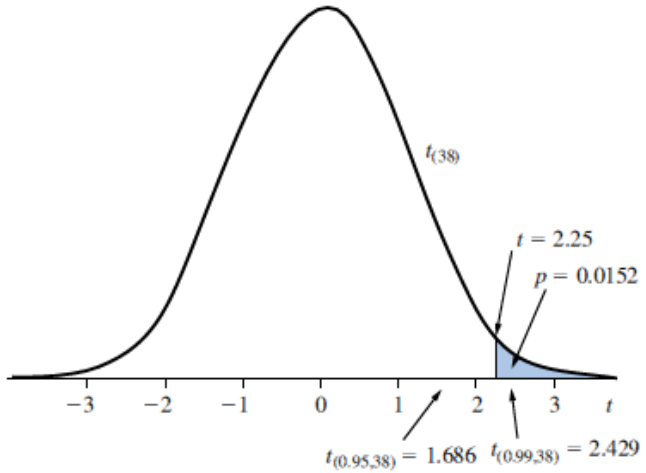
The alternative hypothesis is $H_1$: $\beta_2 > 5.5$

***The t-statistic is***

$$t = \frac{b_2 - 5.5}{\text{se}(b_2)} = \frac{10.21 - 5.5}{2.09} = 2.25$$

***The p-value is***

$$p = P\left[t_{(38)} \geq 2.25\right] = 1 - P\left[t_{(38)} \leq 2.25\right] = 1 - 0.9848 = 0.0152$$



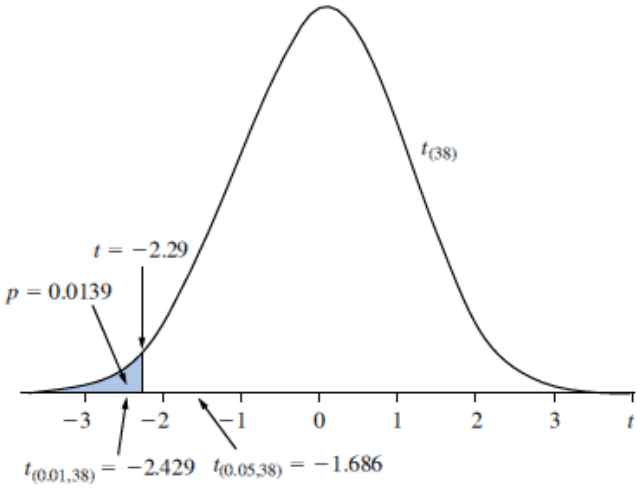
# The p-Value: examples

The null hypothesis is $H_0$: $\beta_2 = 15$

The alternative hypothesis is $H_1$: $\beta_2 < 15$

T-statistic is: $$t = \frac{b_2 - 15}{\text{se}(b_2)} = \frac{10.21 - 15}{2.09} = -2.29$$

The *p*-value is: $$p = P\left[t_{(38)} \leq -2.29\right] = 0.0139$$



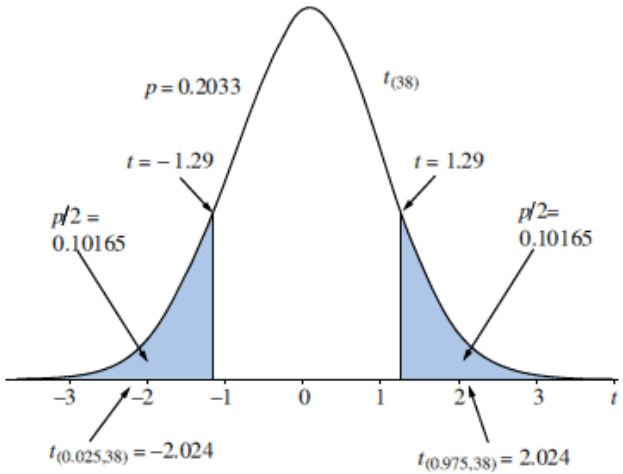
# The p-Value: examples

The null hypothesis is $H_0$: $\beta_2 = 7.5$

The alternative hypothesis is $H_1$: $\beta_2 \neq 7.5$

The t-statistic is: $$t = \frac{b_2 - 7.5}{\text{se}(b_2)} = \frac{10.21 - 7.5}{2.09} = 1.29$$

The *p*-value is: $$p = P\left[t_{(38)} \geq 1.29\right] + P\left[t_{(38)} \leq -1.29\right] = 0.2033$$



Example 5

# The p-Value: examples

The null hypothesis is $H_0$: $\beta_2 = 0$

The alternative hypothesis is $H_1$: $\beta_2 \neq 0$

T-statistic is: $$t = \frac{b_2}{\text{se}(b_2)} = \frac{10.21}{2.09} = 4.88$$

The *p*-value is: $$p = P\left[t_{(38)} \geq 4.88\right] + P\left[t_{(38)} \leq -4.88\right] = 0.0000$$

From typical output, we can easily find the calculated *p*-value used in this example

| Variable | Coefficient | Std. Error | *t*-Statistic | Prob. |
|---|---|---|---|---|
| *C* | 83.41600 | 43.41016 | 1.921578 | 0.0622 |
| *INCOME* | 10.20964 | 2.093264 | 4.877381 | 0.0000 |