POLS0083: Quantitative Data Analysis

Lecture 9: Uncertainty II (Hypothesis Testing)

Julia de Romémont & Michal Ovádek

My bad, I was right!

Confidence interval

A confidence interval is a range of numbers that we believe is likely to *contain* the true difference in means. Confidence intervals are constructed so that they contain the true difference in means in a fixed proportion of samples. This is called the confidence level, which we must select before computing the interval.

- That is **correct**
- What would be **incorrect** is the following:
  - A confidence interval is a range of numbers that we believe are likely to ***be*** the true difference in means.

The "Lady" Tasting Tea (Fisher, 1925)

Imagine you are in an early 1920s agricultural research station.

It is time for tea.....

Dr Bristol Tasting Tea (Fisher, 1925)

- One of the scientists, Dr Muriel Bristol, claims to be able to distinguish whether the milk or the tea had been poured into the cup first.
  - This is a hypothesis

Dr Bristol Tasting Tea (Fisher, 1925)

- One of the scientists, Dr Muriel Bristol, claims to be able to distinguish whether the milk or the tea had been poured into the cup first.
  - This is a hypothesis
- A test was arranged by dubious colleagues.
  - ...the test was 8 cups, 4 of each type, in random order...
  - ...and Dr. Bristol correctly identified all 4 cups into which the milk was poured first.

Dr Bristol Tasting Tea (Fisher, 1925)

- One of the scientists, Dr Muriel Bristol, claims to be able to distinguish whether the milk or the tea had been poured into the cup first.
  - This is a hypothesis
- A test was arranged by dubious colleagues.
  - ...the test was 8 cups, 4 of each type, in random order...
  - ...and Dr. Bristol correctly identified all 4 cups into which the milk was poured first.
- How much evidence is this for Dr Bristol's claim?

Dr Bristol Tasting Tea (Fisher, 1925)

- If Dr Bristol *does not* have the ability to distinguish between milk first and tea first, how likely would it be that she would correctly guess 4 out of 4 tea-first cups?

Dr Bristol Tasting Tea (Fisher, 1925)

- If Dr Bristol *does not* have the ability to distinguish between milk first and tea first, how likely would it be that she would correctly guess 4 out of 4 tea-first cups?
- To figure out the frequency of different possibilities, we ask the following:
  - How many different ways are there to (not) pick 4 cups out of 8?

Successful guesses	Selected Possibilities	Unselected Possibilities	Total Possible Combinations
0	MMMM	TTTT	1×1
1	MMMT, MMTM, MTMM, TMMM	TTTM, TTMT, TMTT, MTTT	4×4
2	MMTT, MTMT, MTTM, TMTM, TTMM, TMMT	TTMM, TMTM, TMMT, MTMT, MMTT, MTTM	6×6
3	MTTT, TMTT, TTMT, TTTM	TMMM, MTMM, MMTM, MMMT	4×4
4	TTTT	MMMM	1×1
		Total	70

What were the chances?

- Assume that Dr Bristol really cannot tell the difference between methods of preparing the tea.
  - This is the null hypothesis

What were the chances?

- Assume that Dr Bristol really cannot tell the difference between methods of preparing the tea.
  - This is the null hypothesis
- How many different ways were there to pick 4 cups out of 8?
  - From the table on the previous slide, we saw that this was 70.

What were the chances?

- Assume that Dr Bristol really cannot tell the difference between methods of preparing the tea.
  - This is the null hypothesis
- How many different ways were there to pick 4 cups out of 8?
  - From the table on the previous slide, we saw that this was 70.
- To perform a *test of the null hypothesis*, we ask:
  - If the null hypothesis is true, what is the probability that we would observe what we observed?

What were the chances?

If the null hypothesis is true, what is the probability that we would observe what we observed?

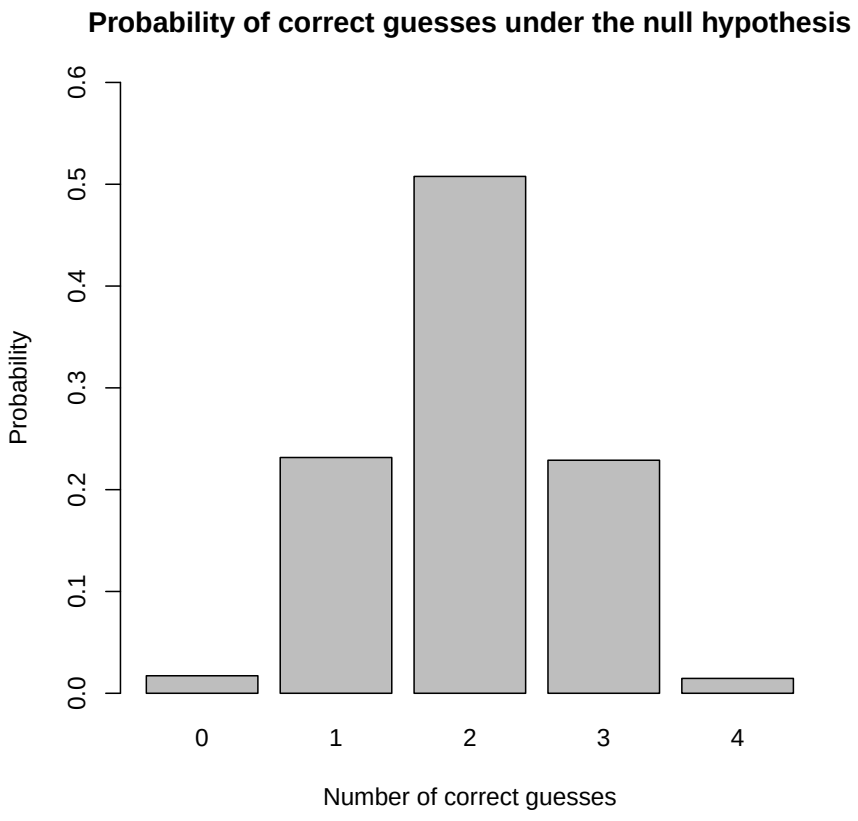
- We observed Dr Bristol correctly selecting 4 out of 4 cups – this is our test statistic

What were the chances?

If the null hypothesis is true, what is the probability that we would observe what we observed?

- We observed Dr Bristol correctly selecting 4 out of 4 cups – this is our test statistic
- We need to derive the sampling distribution of our test statistic **under the assumption that the null hypothesis is true**

What were the chances?

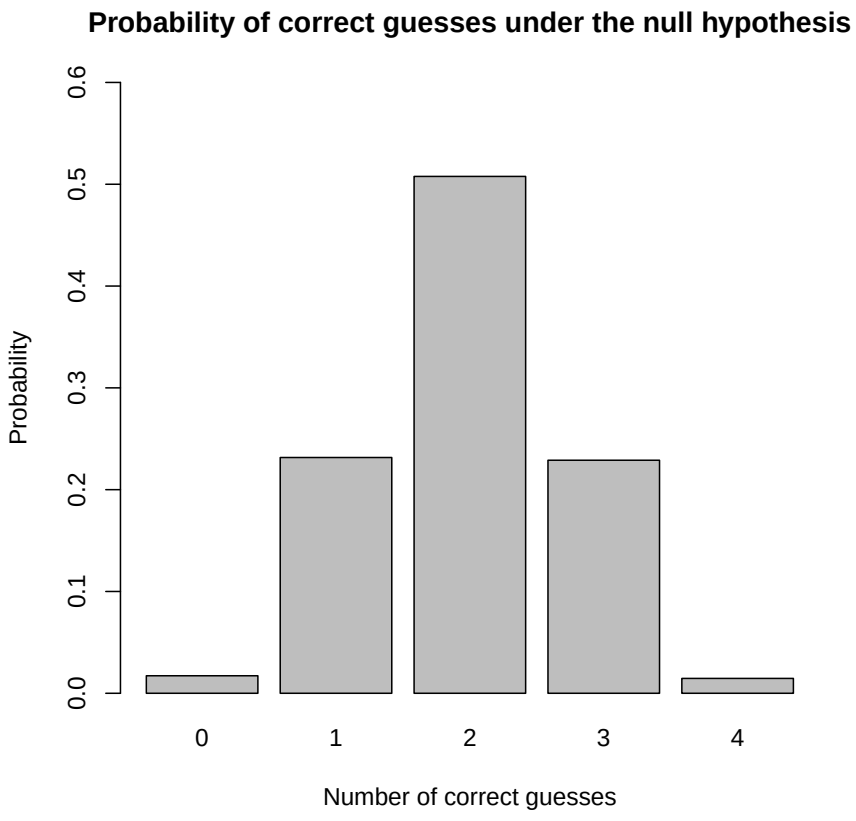


- If Dr Bristol was really only guessing, the probability that she would have correctly identified all four cups of tea:

$$\frac{1}{70} \approx 0.014$$

- This is the p-value

What were the chances?



- If Dr Bristol was really only guessing, the probability that she would have correctly identified all four cups of tea:

$$\frac{1}{70} \approx 0.014$$

- This is the p-value
- It tells us the probability of observing the data we observe under the assumption that the null hypothesis is true.

Does Dr Bristol have special tea-drinking abilities?

What should we conclude?

- It is relatively unlikely that Dr Bristol would have correctly identified the four milk-first cups if she did not have this ability
  - $p = 0.014$
- It is not *impossible* that she simply got lucky...

Does Dr Bristol have special tea-drinking abilities?

What should we conclude?

- It is relatively unlikely that Dr Bristol would have correctly identified the four milk-first cups if she did not have this ability
  - $p = 0.014$
- It is not *impossible* that she simply got lucky...
- Conclusion: We are relatively confident that we can reject the null hypothesis, but we cannot be 100% certain.

Does Dr Bristol have special tea-drinking abilities?

What should we conclude?

- It is relatively unlikely that Dr Bristol would have correctly identified the four milk-first cups if she did not have this ability
  - $p = 0.014$
- It is not *impossible* that she simply got lucky...
- Conclusion: We are relatively confident that we can reject the null hypothesis, but we cannot be 100% certain.

$\Rightarrow$ **The hypothesis tests we will cover today are all based on this type of logic.**

Lecture Outline

Hypothesis Tests

P-values

Confidence intervals

Uncertainty in Regression

Hypothesis Tests in Regression

Conclusion

Hypothesis Tests

Hypothesis testing

The main elements to any hypothesis test are:

1. State the hypothesis and the null hypothesis

Hypothesis testing

The main elements to any hypothesis test are:

1. State the hypothesis and the null hypothesis
2. Calculate a test-statistic

Hypothesis testing

The main elements to any hypothesis test are:

1. State the hypothesis and the null hypothesis
2. Calculate a test-statistic
3. Derive the sampling distribution of the test statistic under the assumption that the null hypothesis is true

Hypothesis testing

The main elements to any hypothesis test are:

1. State the hypothesis and the null hypothesis
2. Calculate a test-statistic
3. Derive the sampling distribution of the test statistic under the assumption that the null hypothesis is true
4. Calculate the p-value

Hypothesis testing

The main elements to any hypothesis test are:

1. State the hypothesis and the null hypothesis
2. Calculate a test-statistic
3. Derive the sampling distribution of the test statistic under the assumption that the null hypothesis is true
4. Calculate the p-value
5. State a conclusion

Hypothesis testing

The main elements to any hypothesis test are:

1. State the hypothesis and the null hypothesis
2. Calculate a test-statistic
3. Derive the sampling distribution of the test statistic under the assumption that the null hypothesis is true
4. Calculate the p-value
5. State a conclusion

When do we reject the null hypothesis?

We reject the null hypothesis if the association we observe between two variables is unlikely to have been observed by chance.

[1] i.e. the null hypothesis were true.

When do we reject the null hypothesis?

We reject the null hypothesis if the association we observe between two variables is unlikely to have been observed by chance.

- This probability is called the α-level and usually takes a value of 0.05 or 0.01.
- When we choose an α-level, we are saying:
  - "I will reject the null hypothesis if the probability that I observed a given relationship (test statistic) in my sample, if in fact there was *no* relationship in the population[1], is below α."

[1] i.e. the null hypothesis were true.

When do we reject the null hypothesis?

We reject the null hypothesis if the association we observe between two variables is unlikely to have been observed by chance.

- This probability is called the α-level and usually takes a value of 0.05 or 0.01.
- When we choose an α-level, we are saying:
  - "I will reject the null hypothesis if the probability that I observed a given relationship (test statistic) in my sample, if in fact there was *no* relationship in the population[1], is below α."
- The confidence level from last week is just $(1 - \alpha) * 100$
  - $\alpha = 0.05$, confidence = 95%
  - $\alpha = 0.01$, confidence = 99%

[1] i.e. the null hypothesis were true.

What does α mean in practice?

- Think back to the Dr Bristol tasting tea example
  - there was a 1.4% probability that Dr Bristol would identify the correct cups just by guessing
  - if we selected $\alpha = 0.05$ we would reject the null hypothesis (that she doesn't have special tea tasting abilities)
  - but she *may* just have been lucky that time![2]

[2] Even if she is just guessing, i.e. doesn't have special tea tasting abilities, she will sometimes get all the cups right by chance and we will incorrectly reject the null.

What does α mean in practice?

- Think back to the Dr Bristol tasting tea example
  - there was a 1.4% probability that Dr Bristol would identify the correct cups just by guessing
  - if we selected $\alpha = 0.05$ we would reject the null hypothesis (that she doesn't have special tea tasting abilities)
  - but she *may* just have been lucky that time![2]
- An α-level of 0.05 implies that, in the process of repeated sampling, we will incorrectly reject the null hypothesis 5% of the time
- An α-level of 0.01 implies that, in the process of repeated sampling, we will incorrectly reject the null hypothesis 1% of the time

[2] Even if she is just guessing, i.e. doesn't have special tea tasting abilities, she will sometimes get all the cups right by chance and we will incorrectly reject the null.

Type-I versus type-II errors

Two potential mistakes of any hypothesis test:

- Type-I error
  - **When we reject a null hypothesis that is true**
  - Or when we find support for a hypothesis that is false
- Type-II error
  - **When we fail to reject a null hypothesis that is false**
  - Or when we do not find support for a hypothesis even though it is true

Type-I versus type-II errors

Two potential mistakes of any hypothesis test:

- Type-I error
  - **When we reject a null hypothesis that is true**
  - Or when we find support for a hypothesis that is false
- Type-II error
  - **When we fail to reject a null hypothesis that is false**
  - Or when we do not find support for a hypothesis even though it is true

There is a trade-off between minimising type-I and type-II errors:

- As we increase our α-level, we are more likely to commit a Type-I error, and less likely to commit a Type-II error

t-tests for the difference in two means

- Often we are interested in whether the mean for one group is different from the mean for another group
  - Do people with health insurance have better health outcomes?
  - Do students in smaller classes get better grades?

t-tests for the difference in two means

- Often we are interested in whether the mean for one group is different from the mean for another group
  - Do people with health insurance have better health outcomes?
  - Do students in smaller classes get better grades?
- A natural hypothesis to test here is whether the means of the two groups are different in the population
- A t-test can be used to conduct a hypothesis test for the difference in means between two groups
- Requires an interval-level dependent variable (Y) and binary independent variable (X)

t-tests for the difference in means

The null hypothesis is that there is no difference between the means of the two groups in the population

$$H_0 : E(Y|X = 1) = E(Y|X = 0) \rightarrow \mu_{Y_{x=1}} - \mu_{Y_{x=0}} = 0$$

t-tests for the difference in means

The null hypothesis is that there is no difference between the means of the two groups in the population

$$H_0 : E(Y|X = 1) = E(Y|X = 0) \rightarrow \mu_{Y_{x=1}} - \mu_{Y_{x=0}} = 0$$

The test statistic for the difference in means (for a null hypothesis of no difference) is

$$t = \frac{(\bar{Y}_{X=1} - \bar{Y}_{X=0}) - (\mu_{Y_{x=1}} - \mu_{Y_{x=0}})}{SE(Y_{X=1} - Y_{X=0})} =$$

t-tests for the difference in means

The null hypothesis is that there is no difference between the means of the two groups in the population

$$H_0 : E(Y|X = 1) = E(Y|X = 0) \rightarrow \mu_{Y_{x=1}} - \mu_{Y_{x=0}} = 0$$

The test statistic for the difference in means (for a null hypothesis of no difference) is

$$t = \frac{(\bar{Y}_{X=1} - \bar{Y}_{X=0}) - (\mu_{Y_{x=1}} - \mu_{Y_{x=0}})}{SE(Y_{X=1} - Y_{X=0})} = \frac{(\bar{Y}_{X=1} - \bar{Y}_{X=0})}{\sqrt{\frac{s^2_{Y_{X=1}}}{n_{X=1}} + \frac{s^2_{Y_{X=0}}}{n_{X=0}}}}$$

- $\mu_{Y_{x=1}} - \mu_{Y_{x=0}}$ is the difference in means under the *null*

t-tests for the difference in means

The null hypothesis is that there is no difference between the means of the two groups in the population

$$H_0 : E(Y|X = 1) = E(Y|X = 0) \rightarrow \mu_{Y_{x=1}} - \mu_{Y_{x=0}} = 0$$

The test statistic for the difference in means (for a null hypothesis of no difference) is

$$t = \frac{(\bar{Y}_{X=1} - \bar{Y}_{X=0}) - (\mu_{Y_{x=1}} - \mu_{Y_{x=0}})}{SE(Y_{X=1} - Y_{X=0})} = \frac{(\bar{Y}_{X=1} - \bar{Y}_{X=0})}{\sqrt{\frac{s^2_{Y_{X=1}}}{n_{X=1}} + \frac{s^2_{Y_{X=0}}}{n_{X=0}}}}$$

- $\mu_{Y_{x=1}} - \mu_{Y_{x=0}}$ is the difference in means under the *null* and is equal to 0!

t-tests for the difference in means

The null hypothesis is that there is no difference between the means of the two groups in the population

$$H_0 : E(Y|X = 1) = E(Y|X = 0) \rightarrow \mu_{Y_{x=1}} - \mu_{Y_{x=0}} = 0$$

The test statistic for the difference in means (for a null hypothesis of no difference) is

$$t = \frac{(\bar{Y}_{X=1} - \bar{Y}_{X=0}) - (\mu_{Y_{x=1}} - \mu_{Y_{x=0}})}{SE(Y_{X=1} - Y_{X=0})} = \frac{(\bar{Y}_{X=1} - \bar{Y}_{X=0})}{\sqrt{\frac{s^2_{Y_{X=1}}}{n_{X=1}} + \frac{s^2_{Y_{X=0}}}{n_{X=0}}}}$$

- $\mu_{Y_{x=1}} - \mu_{Y_{x=0}}$ is the difference in means under the *null* and is equal to 0!
- $s^2_{Y_{X=1}}$ and $s^2_{Y_{X=0}}$ are the sample variances for each group

t-tests for the difference in means

The test statistic for the difference in means is

$$t = \frac{\bar{Y}_{X=1} - \bar{Y}_{X=0} - 0}{SE(Y_{X=1} - Y_{X=0})}$$

t-tests for the difference in means

The test statistic for the difference in means is

$$t = \frac{\bar{Y}_{X=1} - \bar{Y}_{X=0} - 0}{SE(Y_{X=1} - Y_{X=0})}$$

Intuition:

- More evidence of difference in the population when the sample difference in means is larger

t-tests for the difference in means

The test statistic for the difference in means is

$$t = \frac{\bar{Y}_{X=1} - \bar{Y}_{X=0} - 0}{SE(Y_{X=1} - Y_{X=0})}$$

Intuition:

- More evidence of difference in the population when the sample difference in means is larger
- More evidence of difference in the population when the standard error is smaller

t-tests for the difference in means

The test statistic for the difference in means is

$$t = \frac{\bar{Y}_{X=1} - \bar{Y}_{X=0} - 0}{SE(Y_{X=1} - Y_{X=0})}$$

Intuition:

- More evidence of difference in the population when the sample difference in means is larger
- More evidence of difference in the population when the standard error is smaller
- t **measures the number of standard errors separating the mean of one group from the mean of another group**

t-test example: Class sizes and student grades

The STAR Experiment

The STAR project was a randomized experiment designed to test the causal effects of class sizes on learning. Students in Tennessee schools were randomly assigned either to regular sized classes (22-25 students, the **control group**) or to smaller classes (15-17 students, the **treatment group**). We can use the results of this experiment to see the effect of small class sizes on student learning.

- Y (Dependent variable): *grade*
  - Student grade on a standardised math test (0 to 100)
- X (Independent variable): *small_class*
  - `TRUE` = student in small class, `FALSE` = student in regular sized class

Difference-in-means

```r
# Treated mean grade
mean_grade_treated <- mean(star$grade[star$small_class == TRUE])
mean_grade_treated
```

```
## [1] 48.04885
```

```r
# Control mean grade
mean_grade_control <- mean(star$grade[star$small_class == FALSE])
mean_grade_control
```

```
## [1] 45.82546
```

```r
# Difference in means
diff_in_means <- mean_grade_treated - mean_grade_control
diff_in_means
```

```
## [1] 2.22339
```

Difference-in-means

```r
# Treated mean grade
mean_grade_treated <- mean(star$grade[star$small_class == TRUE])
mean_grade_treated
```

```
## [1] 48.04885
```

```r
# Control mean grade
mean_grade_control <- mean(star$grade[star$small_class == FALSE])
mean_grade_control
```

```
## [1] 45.82546
```

```r
# Difference in means
diff_in_means <- mean_grade_treated - mean_grade_control
diff_in_means
```

```
## [1] 2.22339
```

The results suggest that small classes cause modest improvements in student grades.

t-statistic

```r
# Group variance
var_grade_treated <- var(star$grade[star$small_class == 1])
var_grade_control <- var(star$grade[star$small_class == 0])
# Group n
n_treated <- sum(star$small_class == 1)
n_control <- sum(star$small_class == 0)
# standard error
std_error <- sqrt((var_grade_treated/n_treated) +
                    (var_grade_control/n_control))
std_error
```

```
## [1] 0.3486694
```

t-statistic

```r
# Group variance
var_grade_treated <- var(star$grade[star$small_class == 1])
var_grade_control <- var(star$grade[star$small_class == 0])
# Group n
n_treated <- sum(star$small_class == 1)
n_control <- sum(star$small_class == 0)
# standard error
std_error <- sqrt((var_grade_treated/n_treated) +
                    (var_grade_control/n_control))
std_error
```

```
## [1] 0.3486694
```

```r
# t-statistic
t_stat <- diff_in_means/std_error
t_stat
```

```
## [1] 6.376787
```

P-values

Hypothesis testing

The main elements to any hypothesis test are:

1. State the hypothesis and the null hypothesis
2. Calculate a test-statistic
3. Derive the sampling distribution of the test statistic under the assumption that the null hypothesis is true
4. Calculate the p-value
5. State a conclusion

Hypothesis testing

The main elements to any hypothesis test are:

1. State the hypothesis and the null hypothesis
2. Calculate a test-statistic
3. Derive the sampling distribution of the test statistic under the assumption that the null hypothesis is true
4. Calculate the p-value
5. State a conclusion

Sampling distribution for the t-statistic

Just as in the Dr Bristol tasting tea example, we can ask: "what are the chances that we would observe that test statistic if the null hypothesis were true?"

Sampling distribution for the t-statistic

Just as in the Dr Bristol tasting tea example, we can ask: "what are the chances that we would observe that test statistic if the null hypothesis were true?"

- The distribution of our test statistic under the null hypothesis will follow a standard normal distribution because of the **central limit theorem**.

Sampling distribution for the t-statistic

Just as in the Dr Bristol tasting tea example, we can ask: "what are the chances that we would observe that test statistic if the null hypothesis were true?"

- The distribution of our test statistic under the null hypothesis will follow a standard normal distribution because of the **central limit theorem**.
- The t-statistic measures the number of standard errors separating the mean of one group from the mean of another group.

Sampling distribution for the t-statistic

Just as in the Dr Bristol tasting tea example, we can ask: "what are the chances that we would observe that test statistic if the null hypothesis were true?"

- The distribution of our test statistic under the null hypothesis will follow a standard normal distribution because of the **central limit theorem**.
- The t-statistic measures the number of standard errors separating the mean of one group from the mean of another group.
- *If the population difference in means is 0*, 95% of the samples we draw should result in a difference in means that is between 1.96 standard errors from 0

Sampling distribution for the t-statistic

Just as in the Dr Bristol tasting tea example, we can ask: "what are the chances that we would observe that test statistic if the null hypothesis were true?"

- The distribution of our test statistic under the null hypothesis will follow a standard normal distribution because of the **central limit theorem**.
- The t-statistic measures the number of standard errors separating the mean of one group from the mean of another group.
- *If the population difference in means is 0*, 95% of the samples we draw should result in a difference in means that is between 1.96 standard errors from 0

Implication: If our test-statistic is greater than 1.96 or less than -1.96, this suggests it is unlikely that the population difference in means is 0!

P-values for the difference in means

The **probability** of observing our estimated test statistic under the null hypothesis is called the p-value.

P-value

The p-value is the probability that the test-statistic we observe in our sample, or a more extreme value, would be generated in other samples from the population if the null hypothesis was true.

P-values for the difference in means

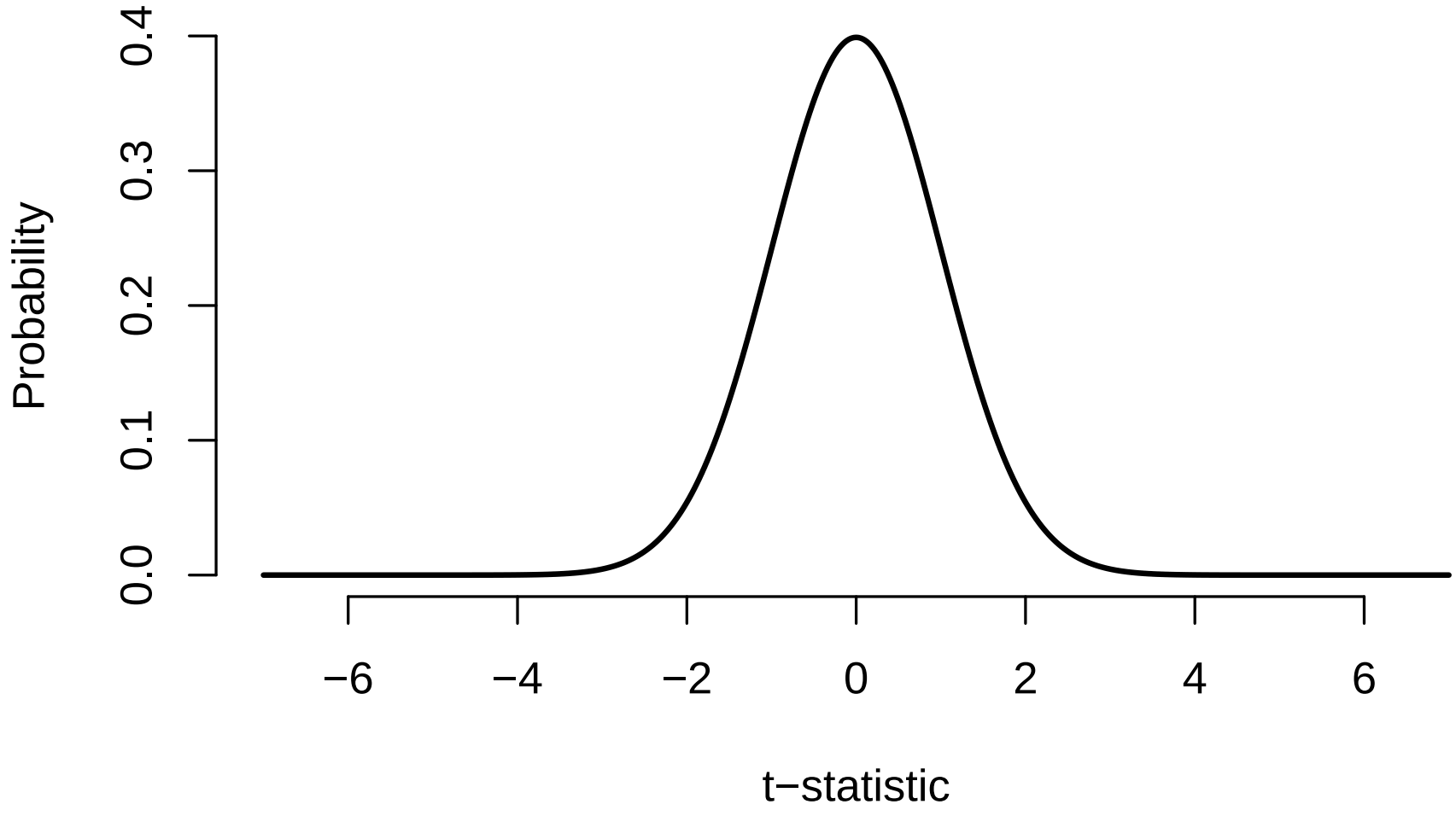
The **probability** of observing our estimated test statistic under the null hypothesis is called the p-value.

P-value

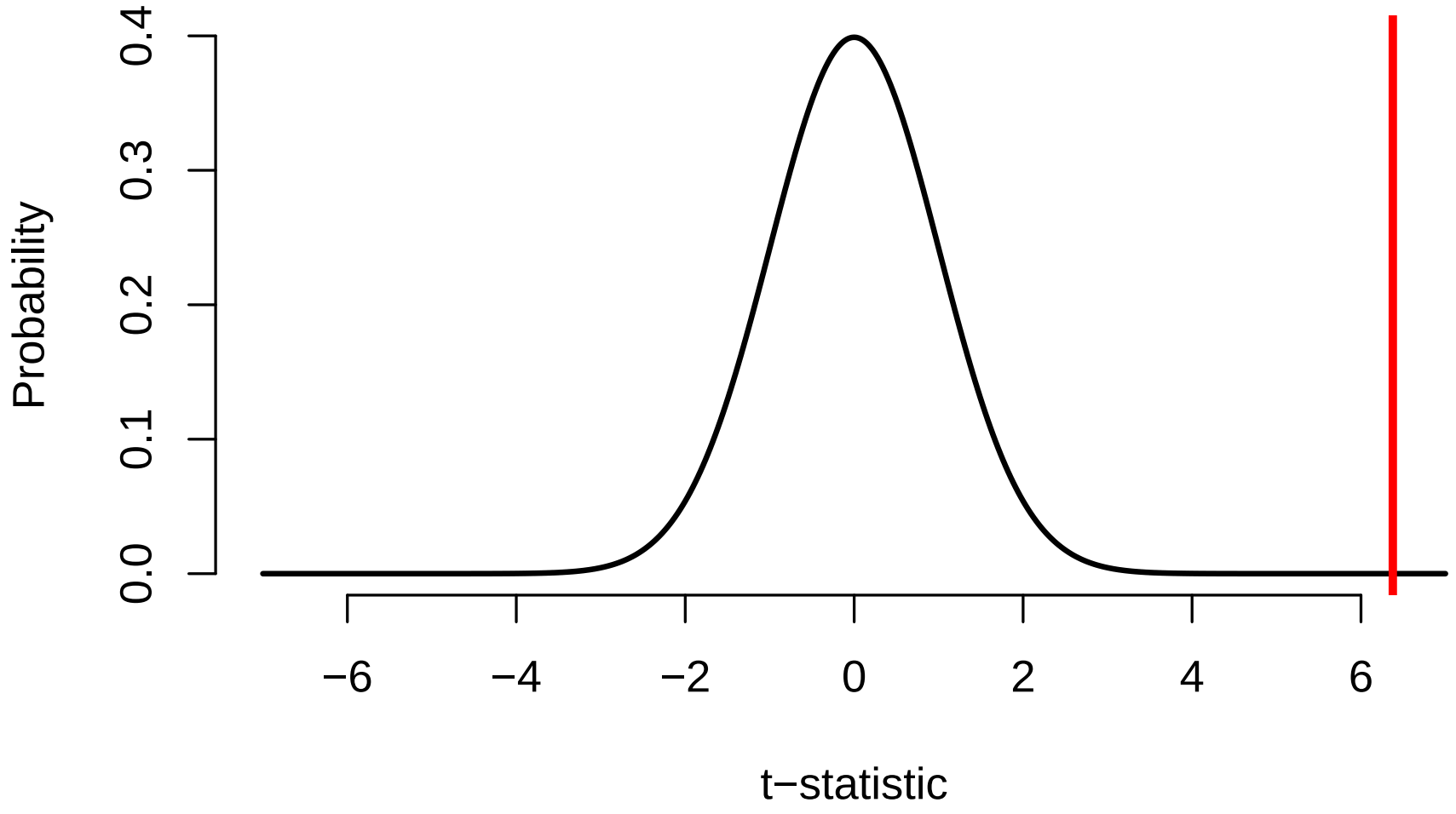
The p-value is the probability that the test-statistic we observe in our sample, or a more extreme value, would be generated in other samples from the population if the null hypothesis was true.

In our case, what is the probability of seeing a difference of means of at least 2.22 **in our sample** if the **population** difference in means is 0?

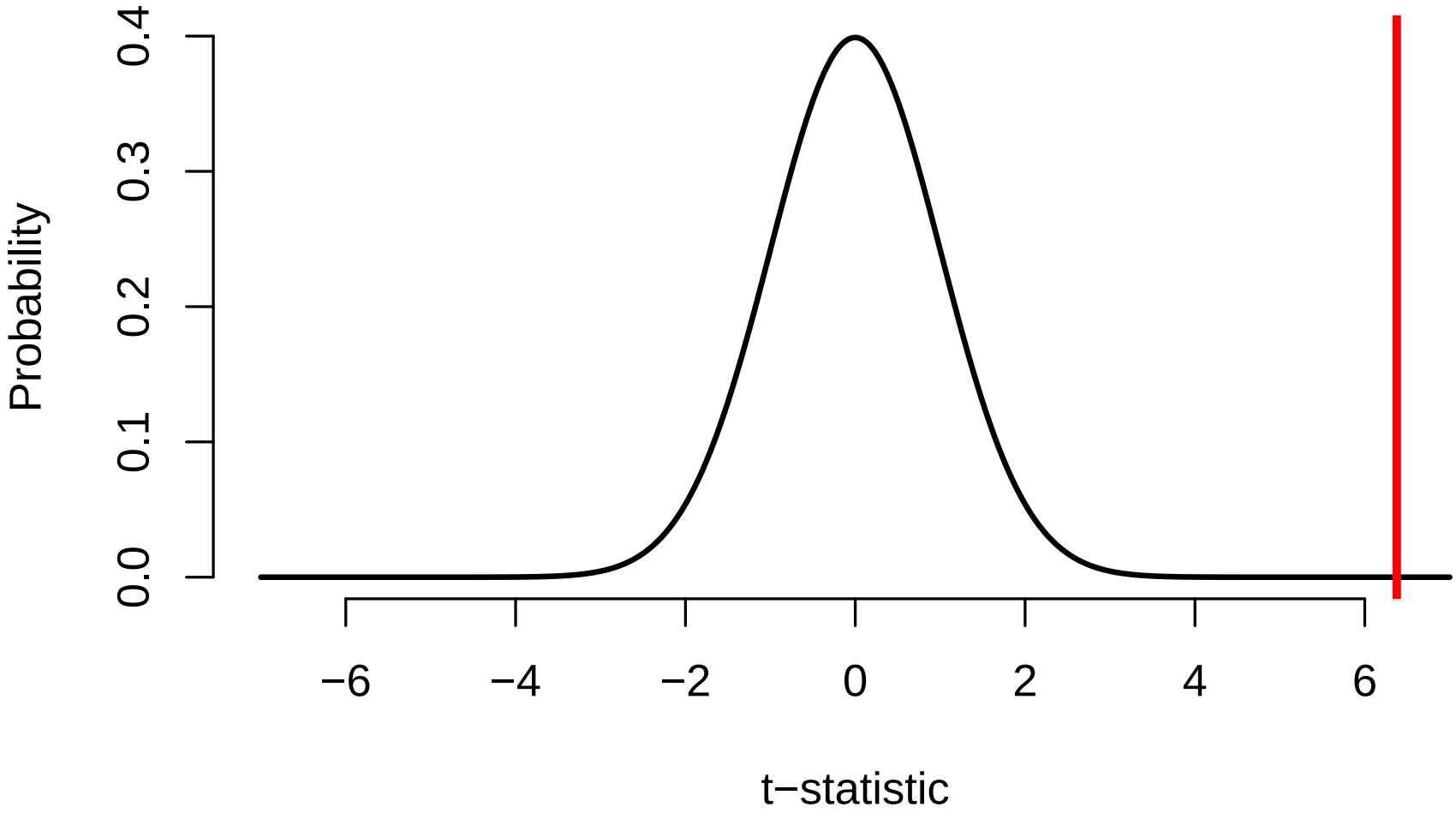
P-values for the difference in means



P-values for the difference in means



P-values for the difference in means



Implication: If the population difference in means is 0, then the probability of observing a test-statistic of 6.38 in any given sample is tiny!

t-test example: Class sizes and student grades

```
t.test(x = star$grade[star$small_class == 1],
       y = star$grade[star$small_class == 0])
```

```
##
##  Welch Two Sample t-test
##
## data:  star$grade[star$small_class == 1] and star$grade[star$small_c
## t = 6.3768, df = 3129.1, p-value = 2.075e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.539747 2.907034
## sample estimates:
## mean of x mean of y
##  48.04885  45.82546
```

Note: 2.075e-10 = 0.000000002074695

t-test example: Class sizes and student grades

```
t.test(x = star$grade[star$small_class == 1],
       y = star$grade[star$small_class == 0])
```

```
##
##  Welch Two Sample t-test
##
## data:  star$grade[star$small_class == 1] and star$grade[star$small_c
## t = 6.3768, df = 3129.1, p-value = 2.075e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.539747 2.907034
## sample estimates:
## mean of x mean of y
##  48.04885  45.82546
```

Note: 2.075e-10 = 0.000000002074695

P-values and t-statistics

Both t-statistics and p-values allow us to assess the amount of evidence we have against the null hypothesis.

Interpretation:

- A *large* t-statistic provides evidence *against* the null hypothesis
  - A t-statistic larger than 1.96 (or smaller than -1.96) allows us to reject the null at the 95% confidence level
  - A t-statistic larger than 2.58 (or smaller than -2.58) allows us to reject the null at the 99% confidence level

P-values and t-statistics

Both t-statistics and p-values allow us to assess the amount of evidence we have against the null hypothesis.

Interpretation:

- A *large* t-statistic provides evidence *against* the null hypothesis
  - A t-statistic larger than 1.96 (or smaller than -1.96) allows us to reject the null at the 95% confidence level
  - A t-statistic larger than 2.58 (or smaller than -2.58) allows us to reject the null at the 99% confidence level
- A *small* p-value provides evidence *against* the null hypothesis
  - A p-value smaller than 0.05 allows us to reject the null at the 95% confidence level
  - A p-value smaller than 0.01 allows us to reject the null at the 99% confidence level

Confidence intervals

From hypothesis tests to confidence intervals

In our class size example we found that we could reject the null hypothesis (at $\alpha = 0.05$) that $\mu_{Y_{X=1}} - \mu_{Y_{X=0}}$ was equal to 0:

$$t = \frac{(48.05 - 45.83) - 0}{0.35} = \frac{2.22}{0.35} \approx 6.38$$

From hypothesis tests to confidence intervals

In our class size example we found that we could reject the null hypothesis (at $\alpha = 0.05$) that $\mu_{Y_{X=1}} - \mu_{Y_{X=0}}$ was equal to 0:

$$t = \frac{(48.05 - 45.83) - 0}{0.35} = \frac{2.22}{0.35} \approx 6.38$$

But what if we had picked a different null hypothesis?

Can we reject the null that $\mu_{Y_{X=1}} - \mu_{Y_{X=0}}$ is equal to 10 in the population?

$$t = \frac{(48.05 - 45.83) - 10}{0.35} = \frac{-7.78}{0.35} \approx -22.3$$

Yes, we can reject the null hypothesis that $\mu_{Y_{X=1}} - \mu_{Y_{X=0}} = 10$ at $\alpha = 0.05$

From hypothesis tests to confidence intervals

In our class size example we found that we could reject the null hypothesis (at $\alpha = 0.05$) that $\mu_{Y_{X=1}} - \mu_{Y_{X=0}}$ was equal to 0:

$$t = \frac{(48.05 - 45.83) - 0}{0.35} = \frac{2.22}{0.35} \approx 6.38$$

But what if we had picked a different null hypothesis?

Can we reject the null that $\mu_{Y_{X=1}} - \mu_{Y_{X=0}}$ is equal to 2 in the population?

$$t = \frac{(48.05 - 45.83) - 2}{0.35} = \frac{0.22}{0.35} \approx 0.64$$

No, we cannot reject the null hypothesis that $\mu_{Y_{X=1}} - \mu_{Y_{X=0}} = 2$ at $\alpha = 0.05$

From hypothesis tests to confidence intervals

We could do this for all possible values of the population difference in means:

- pick a new value for the null hypothesis
- test to see if we can reject the null

Continuing this process would give the set of values for the population difference in means that cannot be rejected[3] at the 95% confidence level.

[3] In other words: the set of values we cannot rule out

From hypothesis tests to confidence intervals

We could do this for all possible values of the population difference in means:

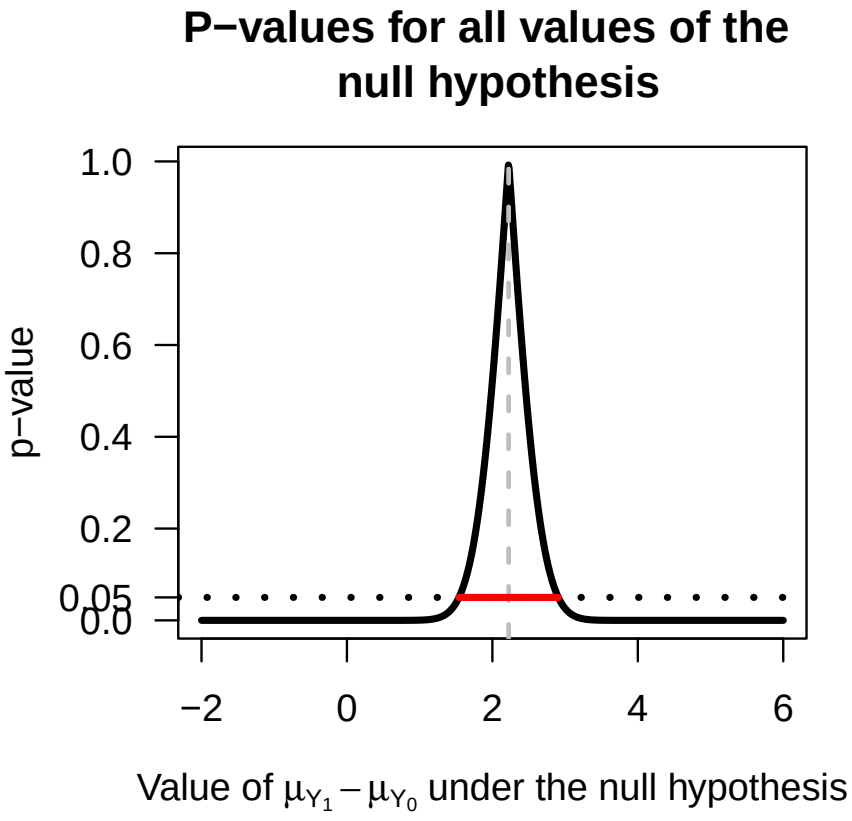
- pick a new value for the null hypothesis
- test to see if we can reject the null

Continuing this process would give the set of values for the population difference in means that cannot be rejected[3] at the 95% confidence level.

This set of values would be our 95% confidence interval!

[3] In other words: the set of values we cannot rule out

From hypothesis tests to confidence intervals



- The red line indicates the range of values for the population difference in means which cannot be rejected at the 95% confidence level.
- This is the 95% confidence interval

α and the confidence level

Note that hypothesis tests and confidence intervals will always give the same result for a given confidence level:

α and the confidence level

Note that hypothesis tests and confidence intervals will always give the same result for a given confidence level:

- If the null hypothesis for the difference in means is 0, and we reject the null hypothesis at the 95% confidence level using a hypothesis test, the value of 0 will not be within the 95% confidence interval!
- If a 95% confidence interval does not include the value of 0,

α and the confidence level

Note that hypothesis tests and confidence intervals will always give the same result for a given confidence level:

- If the null hypothesis for the difference in means is 0, and we reject the null hypothesis at the 95% confidence level using a hypothesis test, the value of 0 will not be within the 95% confidence interval!
- If a 95% confidence interval does not include the value of 0,

α and the confidence level

Note that hypothesis tests and confidence intervals will always give the same result for a given confidence level:

- If the null hypothesis for the difference in means is 0, and we reject the null hypothesis at the 95% confidence level using a hypothesis test, the value of 0 will not be within the 95% confidence interval!
- If a 95% confidence interval does not include the value of 0,

α and the confidence level

Note that hypothesis tests and confidence intervals will always give the same result for a given confidence level:

- If the null hypothesis for the difference in means is 0, and we reject the null hypothesis at the 95% confidence level using a hypothesis test, the value of 0 will not be within the 95% confidence interval!
- If a 95% confidence interval does not include the value of 0, we know that we would also reject the null hypothesis using a hypothesis test with $\alpha = 0.05$.

α and the confidence level

Note that hypothesis tests and confidence intervals will always give the same result for a given confidence level:

```
t.test(x = star$grade[star$small_class == 1],
       y = star$grade[star$small_class == 0])
```

```
##
##  Welch Two Sample t-test
##
## data:  star$grade[star$small_class == 1] and star$grade[star$small_class ==
## t = 6.3768, df = 3129.1, p-value = 2.075e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.539747 2.907034
## sample estimates:
## mean of x mean of y
##  48.04885  45.82546
```

Statistical and "substantive" significance

Our current focus on statistical significance should add to, not replace, our interest in the substantive significance of our results.

- "Statistical" significance is largely a function of sample size.
- If the sample size is very large, the standard error will be small, and so will the p-value.
- But, this does not mean that the relationship is meaningful from a substantive perspective!

Is the effect meaningful?

Imagine that we conduct two new class size experiments.

The first experiment has $N = 20,000$, and we find the following:

- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 0.1$
- $SE(\bar{Y}_{X=1} - \bar{Y}_{X=0}) = 0.02$
- $t = 5$
- $p = 0.0000006$

The second experiment has $N = 1,000$, and we find the following:

- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 4$
- $SE(\bar{Y}_{X=1} - \bar{Y}_{X=0}) = 2$
- $t = 2$
- $p = 0.05$

Is the effect meaningful?

Imagine that we conduct two new class size experiments.

The first experiment has $N = 20,000$, and we find the following:

- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 0.1$
- $SE(\bar{Y}_{X=1} - \bar{Y}_{X=0}) = 0.02$
- $t = 5$
- $p = 0.0000006$

The second experiment has $N = 1,000$, and we find the following:

- $\bar{Y}_{X=1} - \bar{Y}_{X=0} = 4$
- $SE(\bar{Y}_{X=1} - \bar{Y}_{X=0}) = 2$
- $t = 2$
- $p = 0.05$

The first experiment gives a more precise *statistical* result, but the second suggests a more *substantively* important treatment effect.

Uncertainty in Regression

Sampling uncertainty in regression

We just saw that sampling variation means that quantity of interest we calculate will vary across samples.

[4] Important: here we mean the regression intercept, **not** the α-level

Sampling uncertainty in regression

We just saw that sampling variation means that quantity of interest we calculate will vary across samples.

The same applies with regression coefficients – $\hat{\alpha}$[4] and $\hat{\beta}$ – which are also computed from our samples, and therefore are also subject to sampling variation.

We therefore may want to:

[4] Important: here we mean the regression intercept, **not** the α-level

Sampling uncertainty in regression

We just saw that sampling variation means that quantity of interest we calculate will vary across samples.

The same applies with regression coefficients – $\hat{\alpha}$[4] and $\hat{\beta}$ – which are also computed from our samples, and therefore are also subject to sampling variation.

We therefore may want to:

- quantify the sampling uncertainty associated with $\hat{\alpha}$ and $\hat{\beta}$
- construct confidence intervals
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$

[4] Important: here we mean the regression intercept, **not** the α-level

Motivation

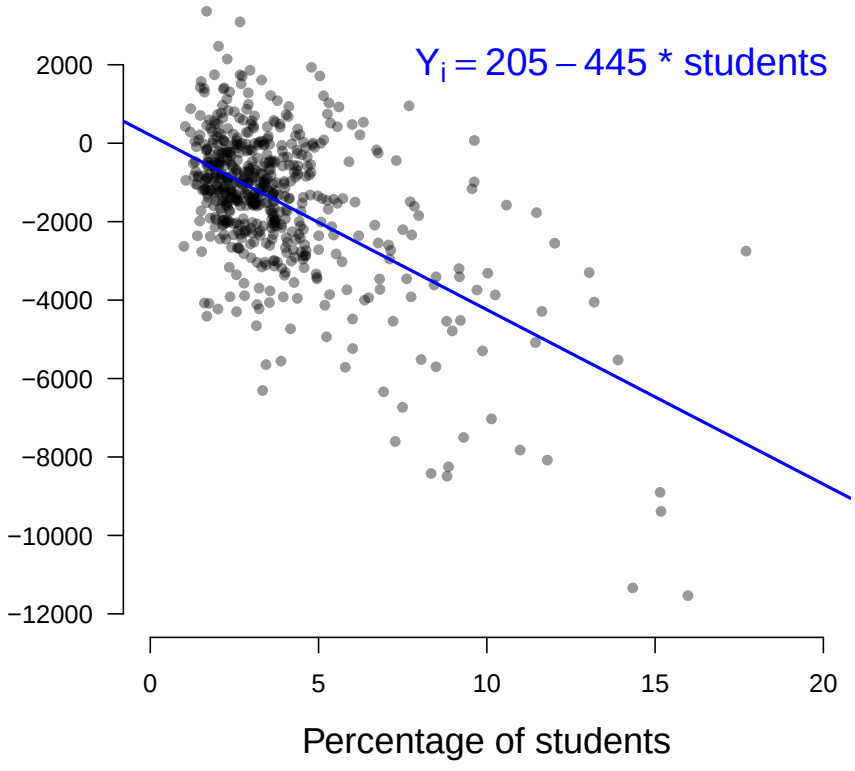
Students and the electoral register

Before 2015 in the UK, the head of the household could register all members of the household to vote. From 2015, all individuals had to register separately. There were particular concerns that this would lead to many students and young people 'falling off' the electoral register. We collect data on voter registration in 573 UK constituencies to evaluate this concern.

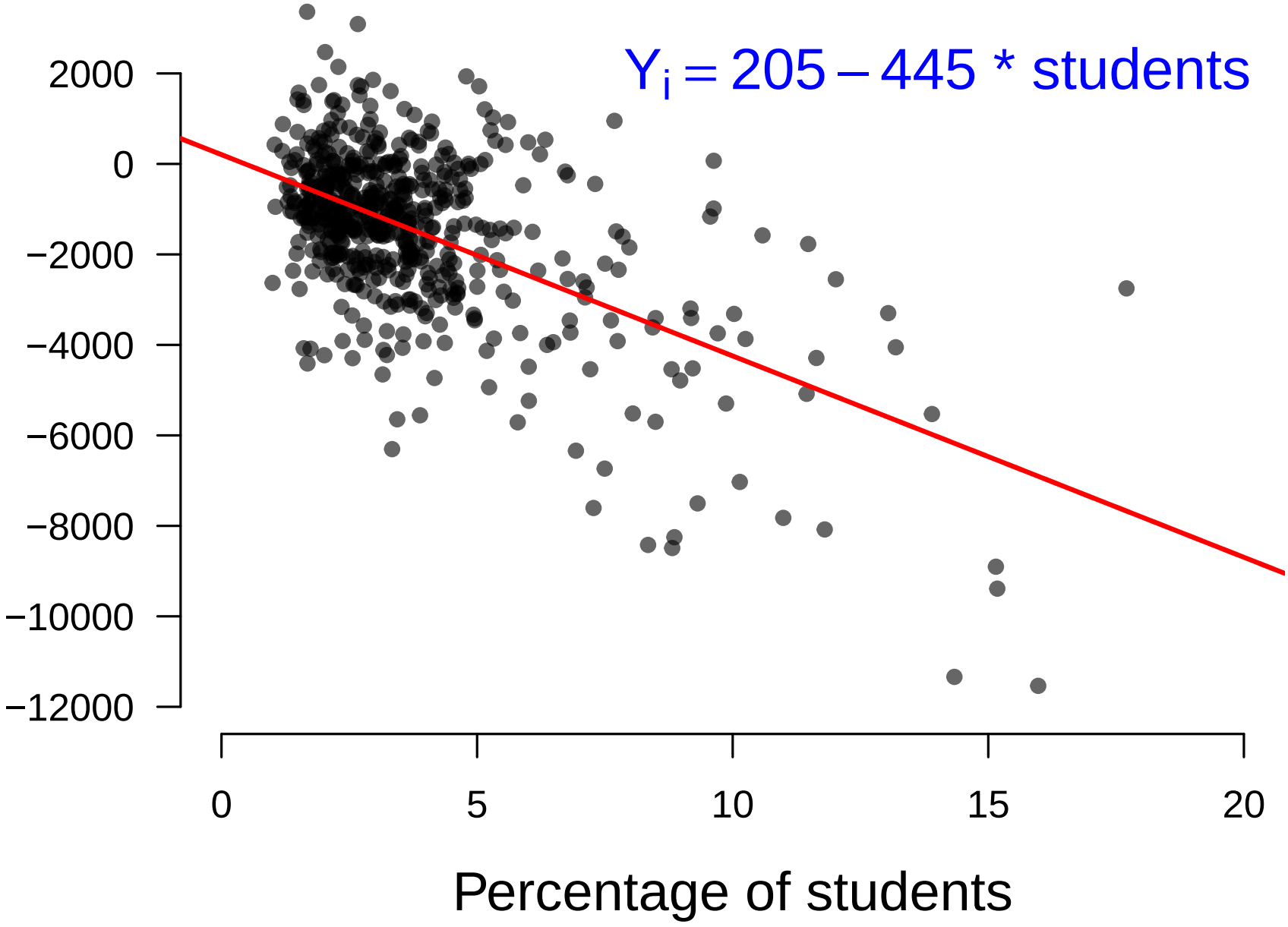
- Unit of analysis: 573 parliamentary constituencies (all constituencies in England and Wales).
- Dependent variable (Y): *Change* in the number of registered voters in a constituency (from 2010 to 2015).
- Independent variable (X): Percentage of a constituency's population who are full time students.

Recap: students and the electoral register

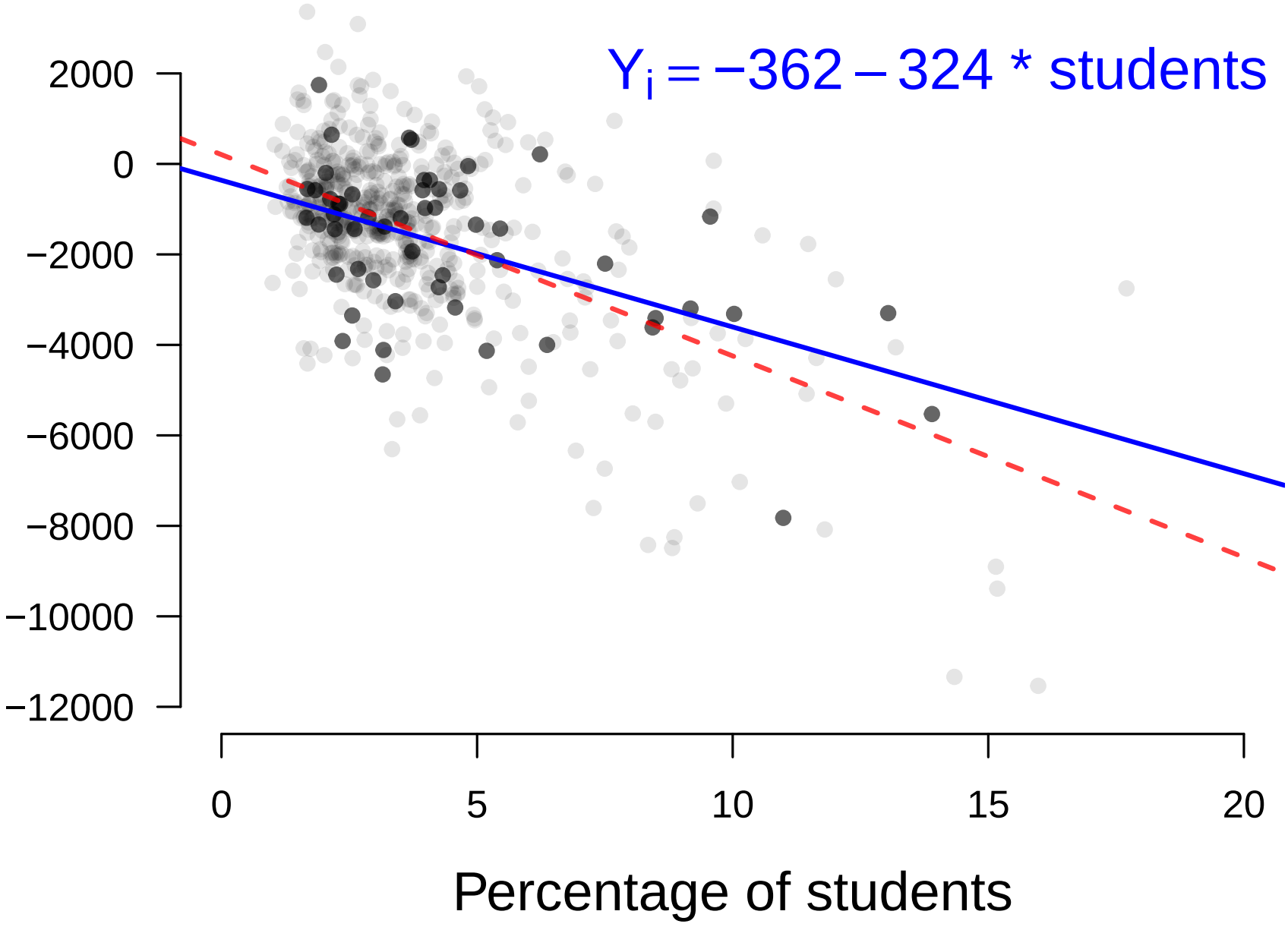
	Change in Reg
students	−444.97***
	(26.99)
Constant	205.15*
	(119.46)
Observations	573
R^2	0.32



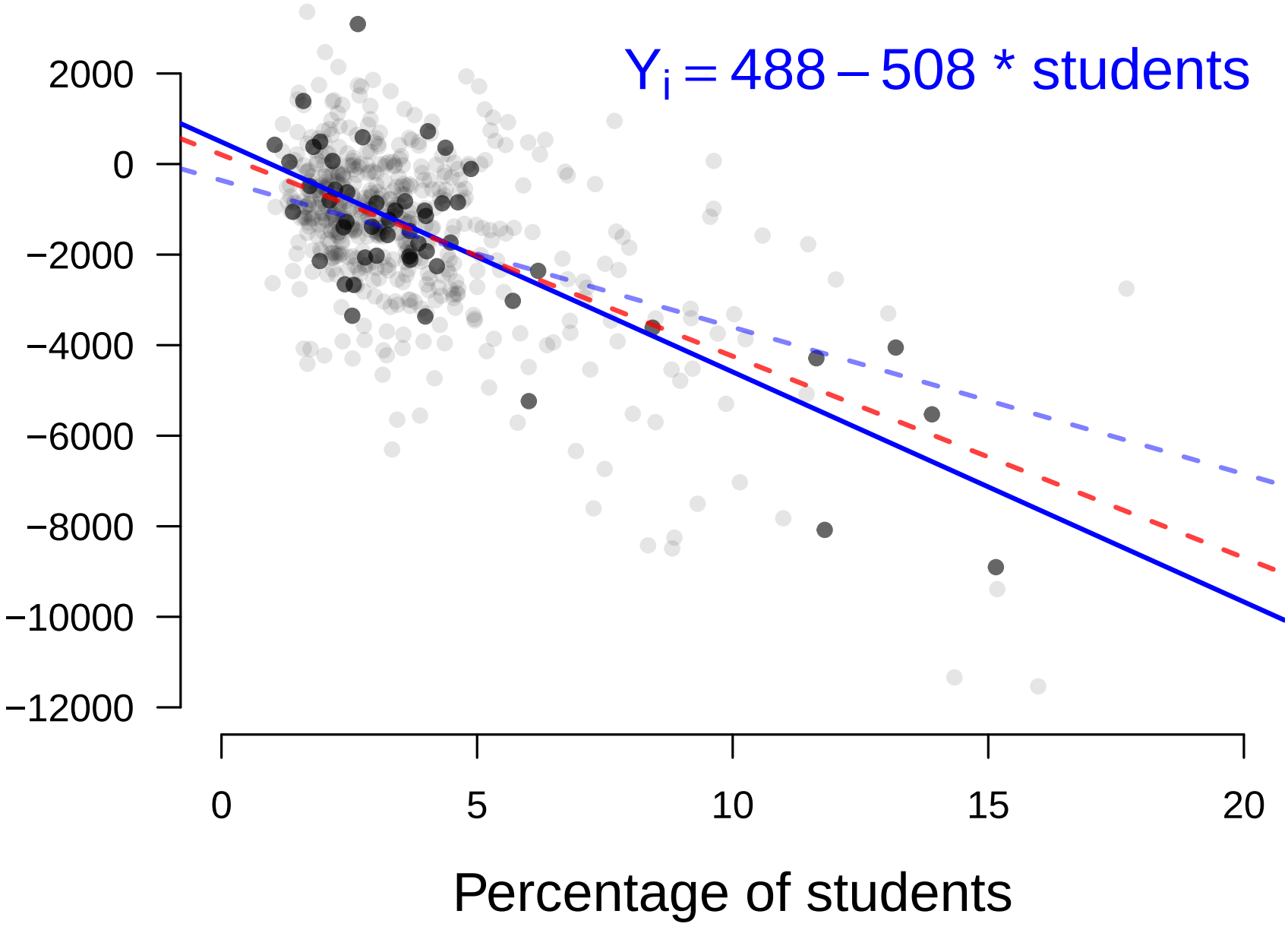
Sampling variation for $\hat{\alpha}$ and $\hat{\beta}$



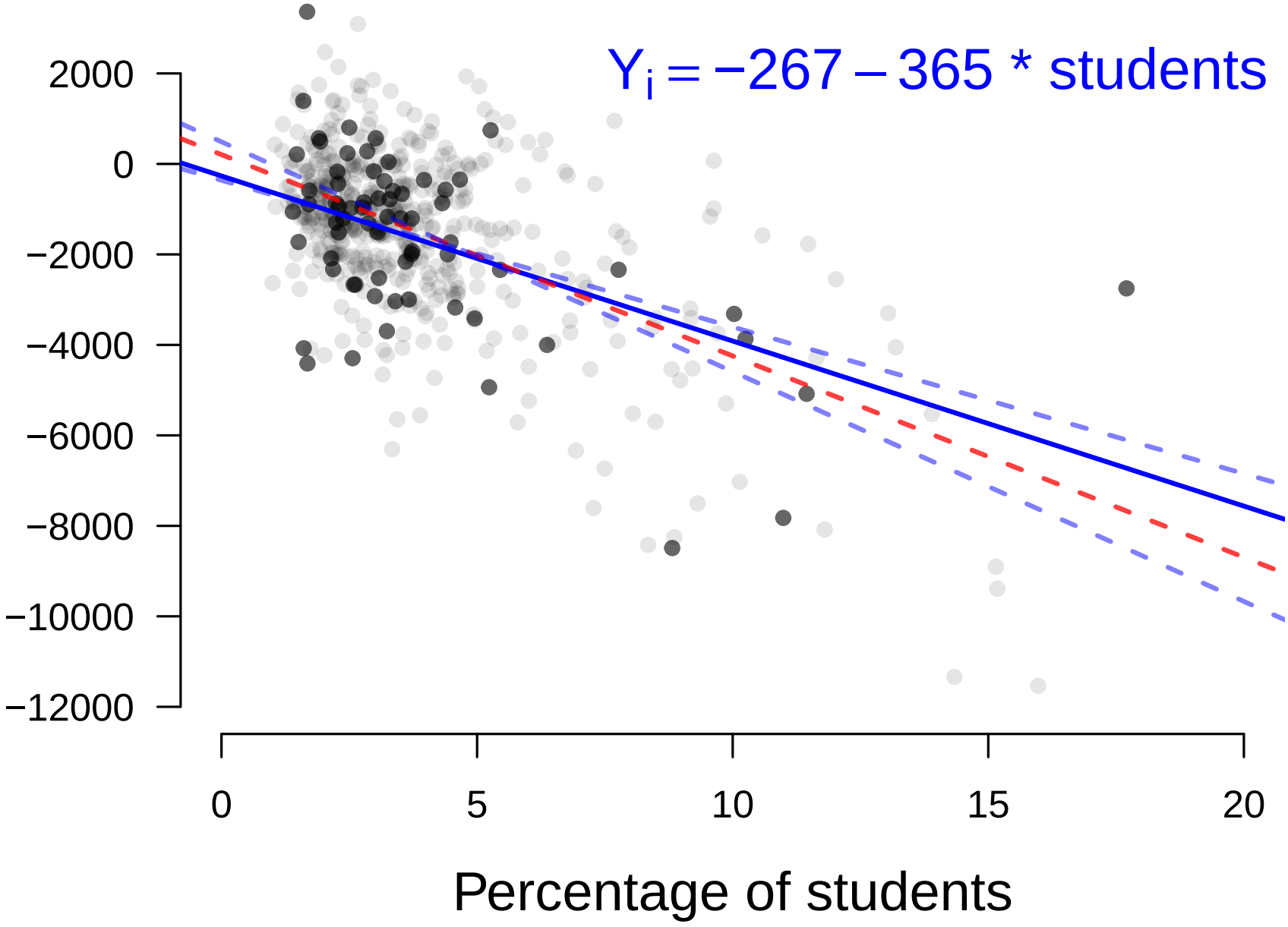
Sampling variation for $\hat{\alpha}$ and $\hat{\beta}$



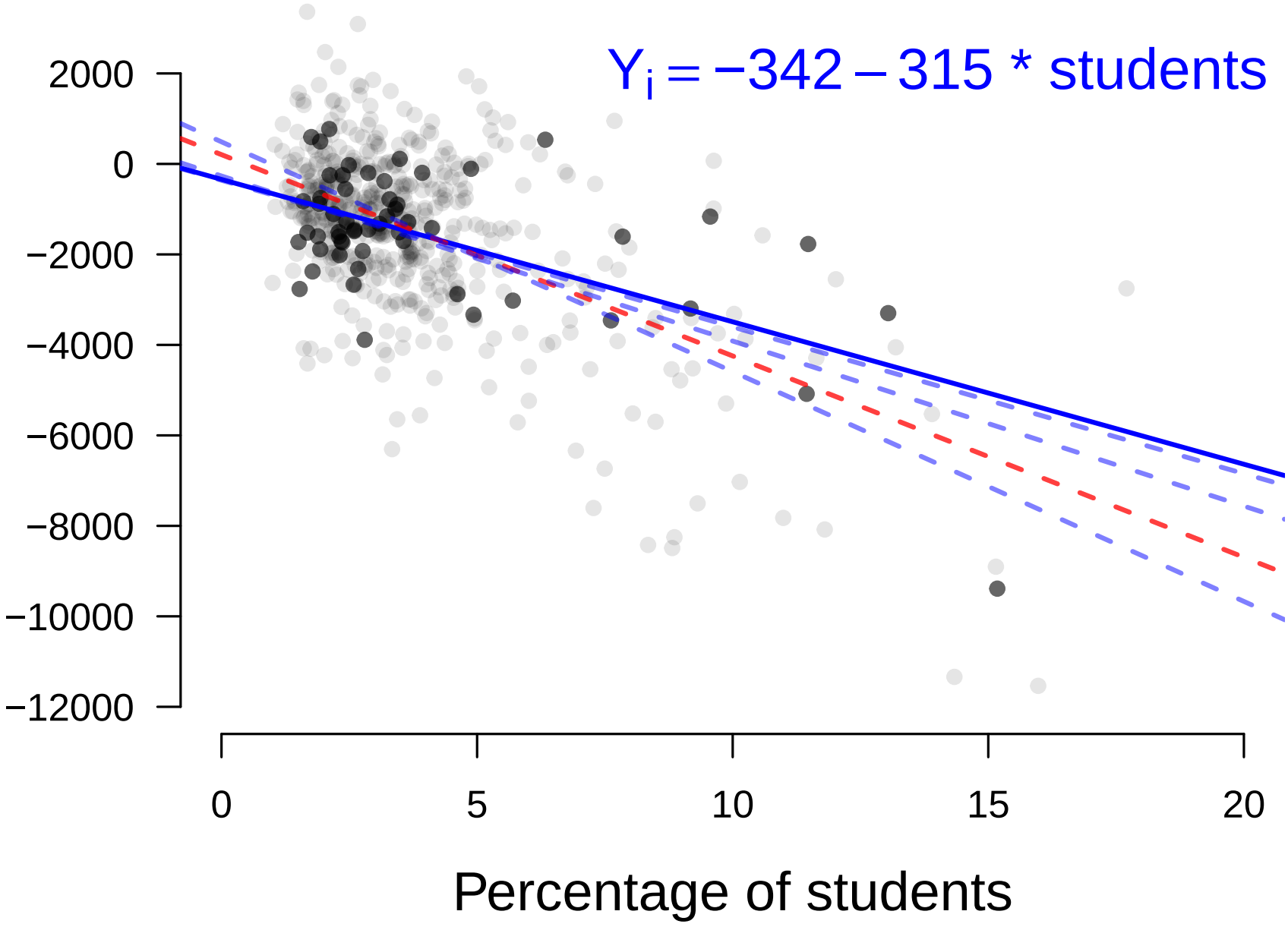
Sampling variation for $\hat{\alpha}$ and $\hat{\beta}$



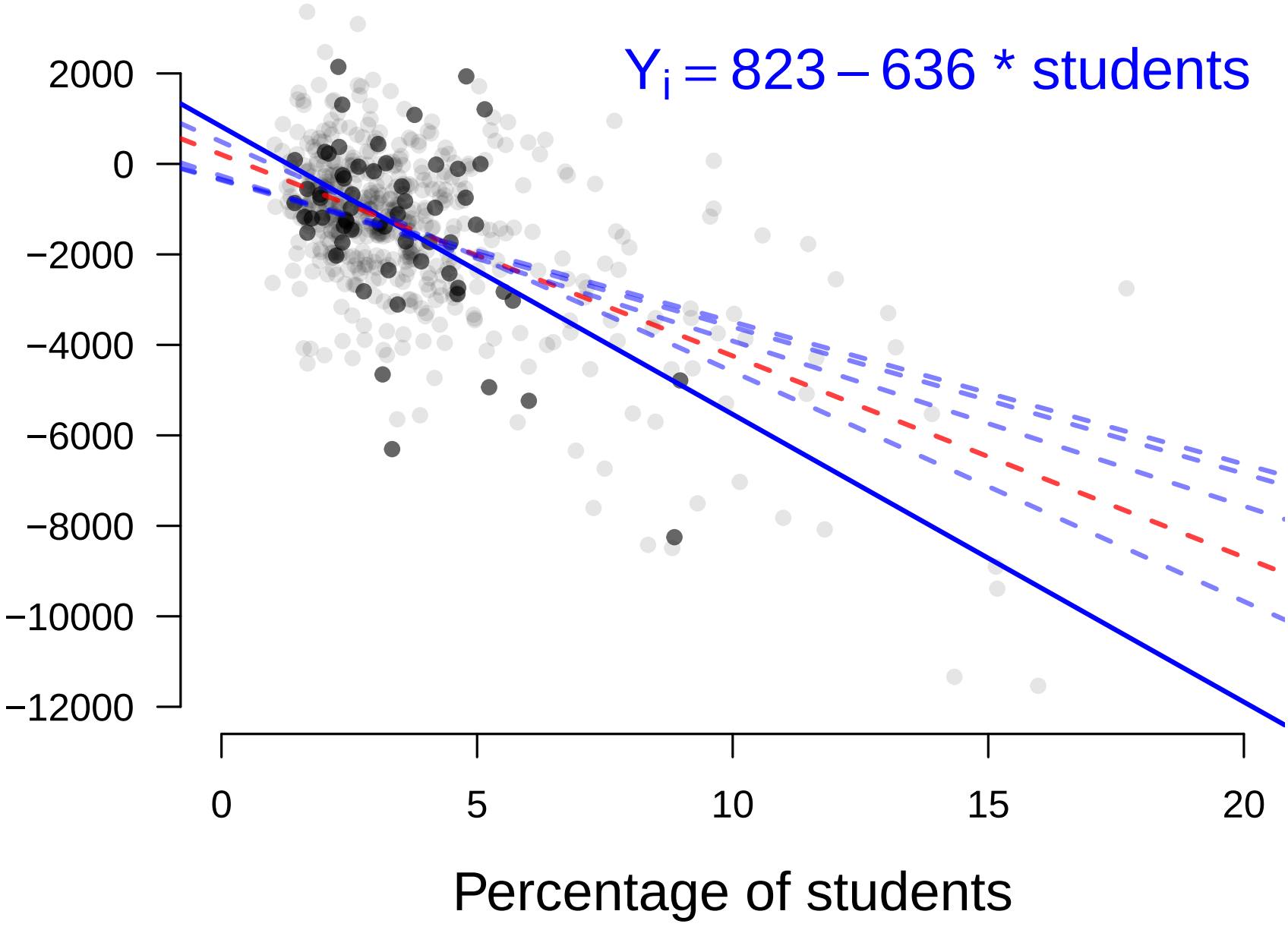
Sampling variation for $\hat{\alpha}$ and $\hat{\beta}$



Sampling variation for $\hat{\alpha}$ and $\hat{\beta}$



Sampling variation for $\hat{\alpha}$ and $\hat{\beta}$



Standard error of regression coefficients

The standard error of the regression coefficients functions in the same way as the standard error of the difference in means:

Standard error of regression coefficients

The standard error of the regression coefficients functions in the same way as the standard error of the difference in means:

- It quantifies the degree of variability we would expect to see across many samples: the sampling distribution
- The Central Limit Theorem tells us that, for non-small-samples, this is all we need to know about the sampling distribution of the regression coefficients, because that sampling distribution will be a normal distribution.

Hypothesis Tests in Regression

In our example

- Problem: The government claimed that the new system of voter registration did not affect students disproportionately.
- In our sample of data, a 1 point increase in the percentage of students in a constituency is associated with, *on average*, a decrease of 445 in the number of registered voters

In our example

- Problem: The government claimed that the new system of voter registration did not affect students disproportionately.
- In our sample of data, a 1 point increase in the percentage of students in a constituency is associated with, *on average*, a decrease of 445 in the number of registered voters
- Is this relationship statistically significantly different from 0?
- How compatible is our estimate with the (null) hypothesis of the government?

Hypothesis tests for regression

Hypothesis tests for regression coefficients are very similar to those for the difference in means:

1. Specify a hypothesis and a null hypothesis
2. Calculate the test-statistic
3. Derive the sampling distribution of the test-statistic under the assumption that the null hypothesis is true
4. Calculate the p-value
5. State a conclusion

Null and alternative hypothesis

In our example:

- H_0: there is no association between the percentage of students and voter registration in the population ($\beta = 0$)
- H_A: there is a non-zero association between the percentage of students and voter registration in the population ($\beta \neq 0$)

Test statistic

The test statistic for a single regression coefficient is:

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}}$$

where $\hat{\sigma}_{\hat{\beta}}$ is the standard error of $\hat{\beta}$.

Test statistic

The test statistic for a single regression coefficient is:

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}}$$

where $\hat{\sigma}_{\hat{\beta}}$ is the standard error of $\hat{\beta}$.

- Note that in the very common case where the null hypothesis is $\beta_{H_0} = 0$ the t-statistic simplifies to $t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$
- You do not need to know how to calculate the standard error ($\hat{\sigma}_{\hat{\beta}}$), but given $\hat{\beta}$ and $\hat{\sigma}_{\hat{\beta}}$, you need to be able to calculate t

The sampling distribution of t

What is the sampling distribution of t?

- When n is large (> 30) the Central Limit Theorem implies that t will follow the standard normal distribution
- When n is small (< 30) the t will follow a t-distribution with n-1 degrees of freedom
- Most regression packages always use the t distribution as the normal distribution is only correct for large sample sizes

The sampling distribution of t

What is the sampling distribution of t?

- When n is large (> 30) the Central Limit Theorem implies that t will follow the standard normal distribution
- When n is small (< 30) the t will follow a t-distribution with n-1 degrees of freedom
- Most regression packages always use the t distribution as the normal distribution is only correct for large sample sizes

Implications:

1. We can normally assume that t will follow the standard normal, unless n is very small
2. We can use the same rules of thumb to assess significance as we did previously (i.e. t > 1.96)

Application to voter registration

For the regression of registration on the percentage of students we obtain:

	Change in Reg
students	-444.97^{***}
	(26.99)
Constant	205.15^{*}
	(119.46)
Observations	573
R^2	0.32

where the numbers in brackets are the standard errors of the coefficients.

Application to voter registration

	Change in Reg
students	-444.97^{***}
	(26.99)
Constant	205.15^{*}
	(119.46)
Observations	573
R^2	0.32

To test the government's hypothesis:

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}} = \frac{-445 - 0}{27} \approx -16$$

Application to voter registration

	Change in Reg
students	-444.97^{***}
	(26.99)
Constant	205.15^{*}
	(119.46)
Observations	573
R^2	0.32

To test the government's hypothesis:

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}} = \frac{-445 - 0}{27} \approx -16$$

Can we reject the null hypothesis at $\alpha = 0.05$?

Application to voter registration

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}} = \frac{-445 - 0}{27} \approx -16$$

- The probability of observing a value of the t-statistic outside the interval $[-1.96, 1.96]$ is less than five percent under the standard normal distribution.

Application to voter registration

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}} = \frac{-445 - 0}{27} \approx -16$$

- The probability of observing a value of the t-statistic outside the interval $[-1.96, 1.96]$ is less than five percent under the standard normal distribution.
- As the t-statistic is clearly outside this interval, the probability that H_0 is correct is less than five percent.
- We can therefore reject the government's claim at the 95% confidence level.

Application to voter registration

R will automatically calculate the correct test-statistic for you:

```
summary(simple_ols_model)
```

```
...
##     Min      1Q  Median      3Q     Max
## -5163.4  -787.0   -21.7   924.5  4921.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   205.15     119.46   1.717   0.0865 .
## students     -444.97      26.99 -16.489   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1525 on 571 degrees of freedom
## Multiple R-squared:  0.3226, Adjusted R-squared:  0.3214
## F-statistic: 271.9 on 1 and 571 DF,  p-value: < 2.2e-16
...
```

Statistical significance

- In the vast majority of t-tests the null hypothesis is that the coefficient is equal to zero.
- In this case the null hypothesis is often not even stated and you will encounter statements such as:
  - The coefficient is statistically significant at the $XX\%$ level
  - The coefficient is statistically significant at conventional levels

Statistical significance

- In the vast majority of t-tests the null hypothesis is that the coefficient is equal to zero.
- In this case the null hypothesis is often not even stated and you will encounter statements such as:
  - The coefficient is statistically significant at the $XX\%$ level
  - The coefficient is statistically significant at conventional levels
- In all of these statements the implicit null hypothesis is that the coefficient of interest is equal to zero.

Statistical significance

- We should not forget that t-test can nevertheless be used to test also other null hypotheses.
- For example, can we reject the null that the true association between the percentage of students and voter registration is -460?

$$t = \frac{\hat{\beta} - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}}} = \frac{-445 - (-460)}{27} \approx 0.56$$

- As 0.56 falls within the interval (-1.96, 1.96), we fail to reject the new null hypothesis that $\beta_{H_0} = -460$

P-values

We can also determine precisely how unlikely the government's hypothesis is given our estimates by calculating the p-value

```
summary(simple_ols_model)
```

```
...
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   205.15     119.46   1.717   0.0865 .
## students     -444.97      26.99 -16.489   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

What does a p-value of $<$ 2e-16 mean?

P-values

We can also determine precisely how unlikely the government's hypothesis is given our estimates by calculating the p-value

```
summary(simple_ols_model)
```

```
...
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   205.15     119.46   1.717   0.0865 .
## students     -444.97      26.99 -16.489   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

What does a p-value of $<$ 2e-16 mean?

- 2e-16 $= 0.0000000000000002$
- $\rightarrow$ it is *very* unlikely that we would observe this test-statistic if the null hypothesis were true

Confidence intervals for regression coefficients

We can also estimate confidence intervals for $\hat{\beta}$:

95% Confidence interval : $\hat{\beta} \pm 1.96 * SE(\hat{\beta})$

99% Confidence interval : $\hat{\beta} \pm 2.58 * SE(\hat{\beta})$

Confidence intervals for regression coefficients

We can also estimate confidence intervals for $\hat{\beta}$:

$$\text{95\% Confidence interval} : \hat{\beta} \pm 1.96 * SE(\hat{\beta})$$

$$\text{99\% Confidence interval} : \hat{\beta} \pm 2.58 * SE(\hat{\beta})$$

In the case of our regression the 95 percent confidence interval:

$$\text{Lower bound: } = -445 - 1.96 \times 27 = -498$$

$$\text{Upper bound: } = -445 + 1.96 \times 27 = -392$$

Confidence intervals for regression coefficients

We can also estimate confidence intervals for $\hat{\beta}$:

$$95\% \text{ Confidence interval} : \hat{\beta} \pm 1.96 * SE(\hat{\beta})$$

$$99\% \text{ Confidence interval} : \hat{\beta} \pm 2.58 * SE(\hat{\beta})$$

In the case of our regression the 95 percent confidence interval:

$$\text{Lower bound: } = -445 - 1.96 \times 27 = -498$$

$$\text{Upper bound: } = -445 + 1.96 \times 27 = -392$$

Intuition: The confidence interval contains all values of the population parameter that cannot be rejected at the five percent significance level given our estimate.

Uncertainty in multiple linear regression

All of the interpretation we have covered for the simple linear regression model translates to more complex linear models:

- Standard errors
- T-statistics
- P-values

Uncertainty in multiple regression: example

```
multiple_ols_model <- lm(voters_change ~ students + urban + wales,
                         data = constituencies)
```

	Change in Reg	
	(1)	(2)
students	−444.97***	−418.21***
	(26.99)	(26.75)
urban		−692.62***
		(126.54)
wales		−345.18
		(243.83)
Constant	205.15*	456.63***
	(119.46)	(124.86)
Observations	573	573
R^2	0.32	0.36

Hypothesis tests in multiple regression: example

	Change in Reg	
	(1)	(2)
students	−444.97***	−418.21***
	(26.99)	(26.75)
urban		−692.62***
		(126.54)
wales		−345.18
		(243.83)
Constant	205.15*	456.63***
	(119.46)	(124.86)
Observations	573	573
R^2	0.32	0.36

t-statistic for $\hat{\beta}_2$ (urban):

$$t = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{-693}{127} \approx -5$$

Hypothesis tests in multiple regression: example

	Change in Reg	
	(1)	(2)
students	-444.97^{***}	-418.21^{***}
	(26.99)	(26.75)
urban		-692.62^{***}
		(126.54)
wales		-345.18
		(243.83)
Constant	205.15^{*}	456.63^{***}
	(119.46)	(124.86)
Observations	573	573
R^2	0.32	0.36

t-statistic for $\hat{\beta}_2$ (urban):

$$t = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{-693}{127} \approx -5$$

t-statistic for $\hat{\beta}_3$ (Wales):

$$t = \frac{\hat{\beta}_3}{\hat{\sigma}_{\hat{\beta}_3}} = \frac{-345}{244} \approx -1$$

Hypothesis tests in multiple regression: example

	Change in Reg	
	(1)	(2)
students	−444.97***	−418.21***
	(26.99)	(26.75)
urban		−692.62***
		(126.54)
wales		−345.18
		(243.83)
Constant	205.15*	456.63***
	(119.46)	(124.86)
Observations	573	573
R^2	0.32	0.36

t-statistic for $\hat{\beta}_2$ (urban):

$$t = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{-693}{127} \approx -5$$

t-statistic for $\hat{\beta}_3$ (Wales):

$$t = \frac{\hat{\beta}_3}{\hat{\sigma}_{\hat{\beta}_3}} = \frac{-345}{244} \approx -1$$

Can we reject the nulls of $\beta_2 = 0$ and $\beta_3 = 0$ at the 95% confidence level, respectively?

Stargazing

You will have noticed that stars (***) are often presented in regression output next to some coefficients.

These are convenient ways to indicate levels of statistical significance. Normally, they represent:

- . = significant at 90% confidence level
- * = significant at 95% confidence level
- ** = significant at 99% confidence level
- *** = significant at 99.9% confidence level

While these are useful, for the assessment you will need to be able to interpret the t-statistics and p-values, not just look for stars in the tables!

Conclusion

What have we learned today?

1. Hypothesis tests can be used to evaluate the plausiblity of a population parameter taking a certain value, given the data we observe in our sample
2. Hypothesis tests and confidence intervals will always give the same conclusion for the same confidence level
3. Regression estimates are computed from samples and so are also subject to sampling variation
4. We need to consider regression standard errors, p-values, and confidence intervals in our interpretations

Seminar

In seminars this week, you will learn to ...

1. Conduct hypothesis tests
2. Construct confidence intervals

Seminar next week

- For next week, there is no 'worksheet' for the seminar as there was before
- Instead, we will ask you to analyse data on the Brexit vote in groups and present your findings
  - More details in the seminars
- There is a homework sheet, but we will immediately release the solutions

Midterm

- Grades should be released at one point during the day
- We will also release the correct solutions
- Use this for revision!
  - In particular, pay attention to how we have written/formatted the code
  - And how we wrote the answer sentence (these tell you how we want youto write the sentences in the final!)