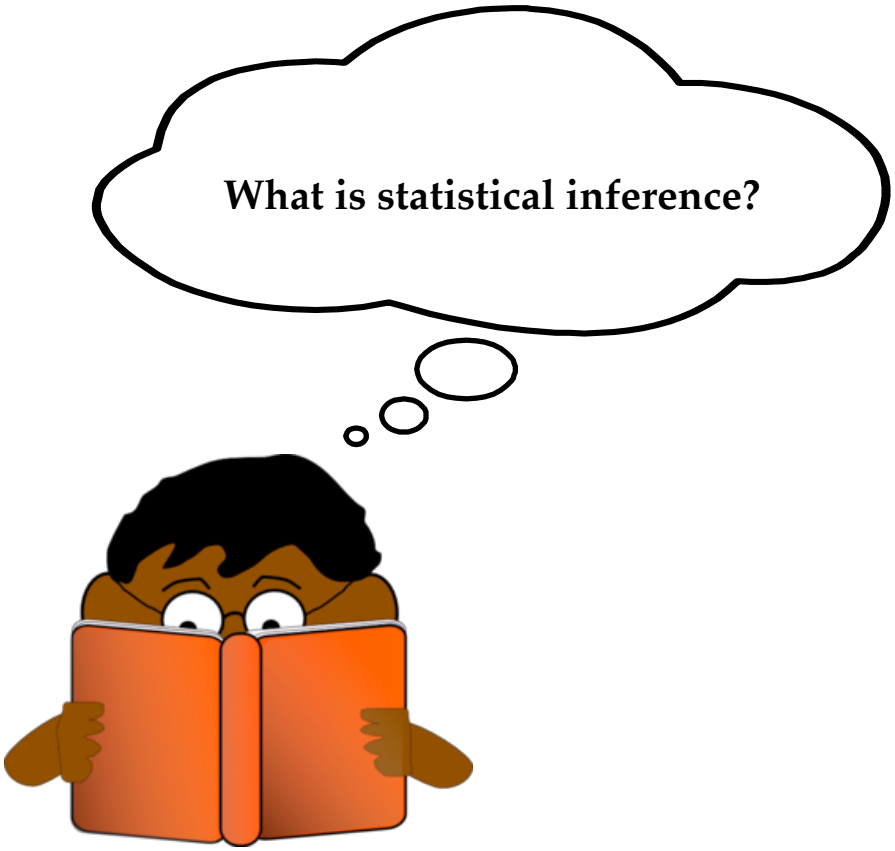
Inference for Simple Regression Models

Prepare

```
# Load the data (homework-achievement.csv)
> math = read.csv("EPSY-8262/data/homework-achievement.csv")

# Load libraries; Note: you may need to install them first
> library(arm)
> library(ggplot2)

# Fit the regression model
> lm.a = lm(achievement ~ homework, data = math)
```



"Statistical inference is used to learn from incomplete or imperfect data." – Gelman & Hill (2006, p. 16)

- **Sampling model:** The primary interest is to learn about one or more characteristics about a population. These characteristics must be estimated from sample data.

- **Measurement error model:** The primary interest is in learning about some underlying pattern or law (maybe to test a theory), but the data are measured with error.

Despite these being very different paradigms, in practice they are often combined (e.g., we measure imperfectly *and* we want to make generalizations)

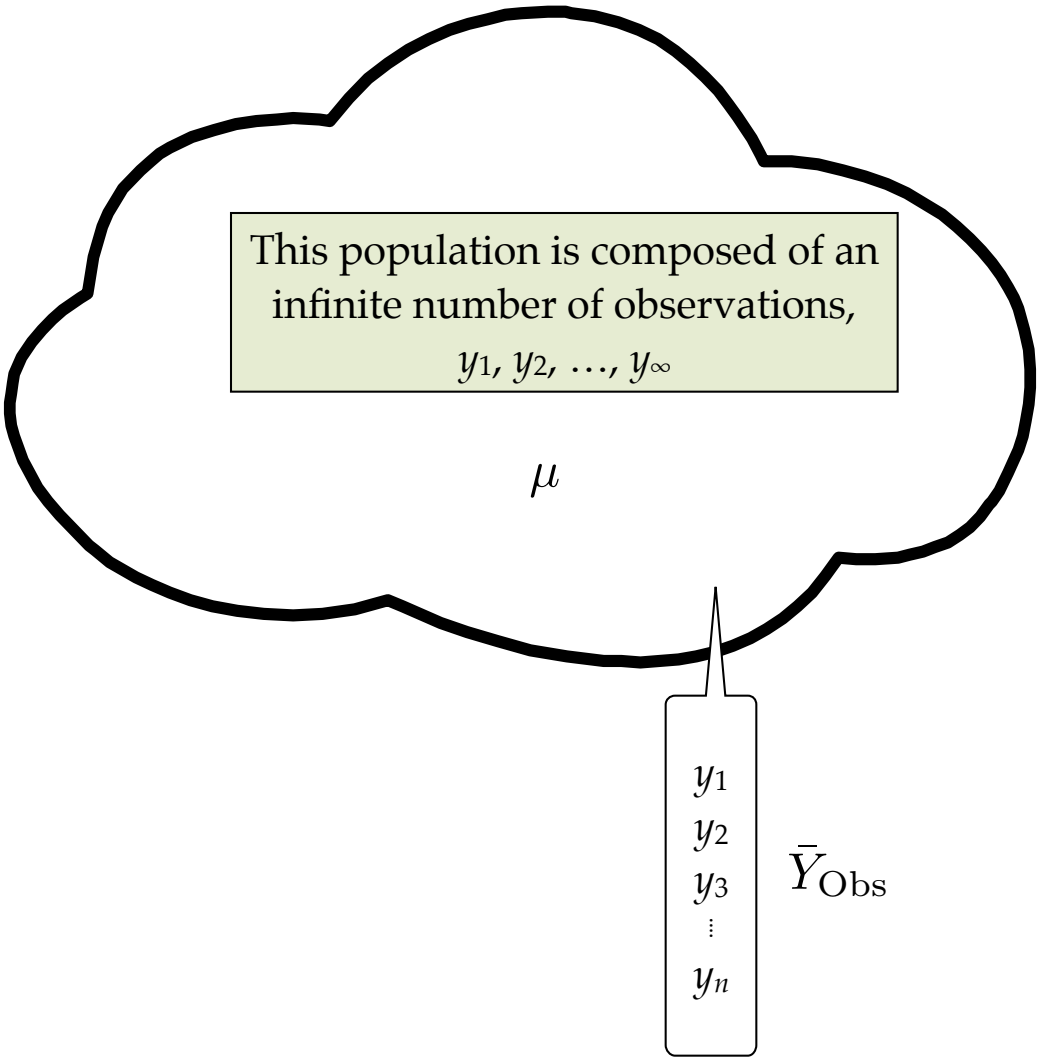
Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel hierarchical models*. New York: Cambridge University Press.

Goals of Statistical Inference

There are many goals of statistical inference, however for conventional regression analysis the goals are: (1) to **estimate the parameters** of the proposed model; and (2) to summarize the amount of **uncertainty** in those estimates.

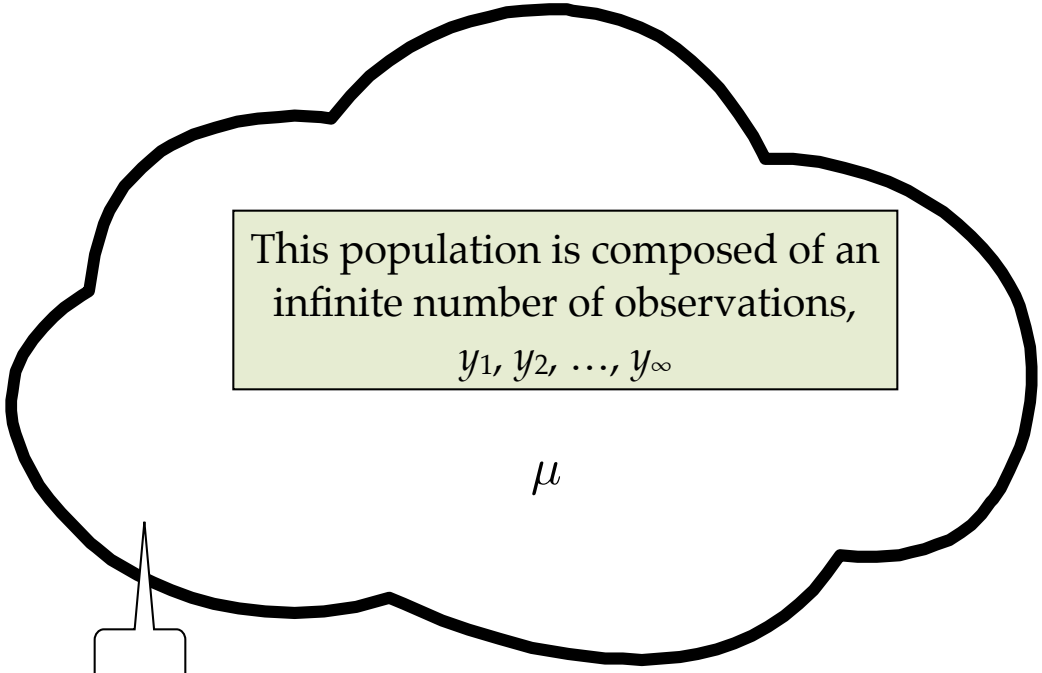
Estimation and Quantification of Uncertainty

Example from Stat I



Goal 1: Estimate the mean for a given population

$$\bar{Y}_{\text{Obs}} \approx \mu$$



y_1
y_2
y_3
$\vdots$
y_n

$\bar{Y}_1$

$\bar{Y}_1 \approx \mu$

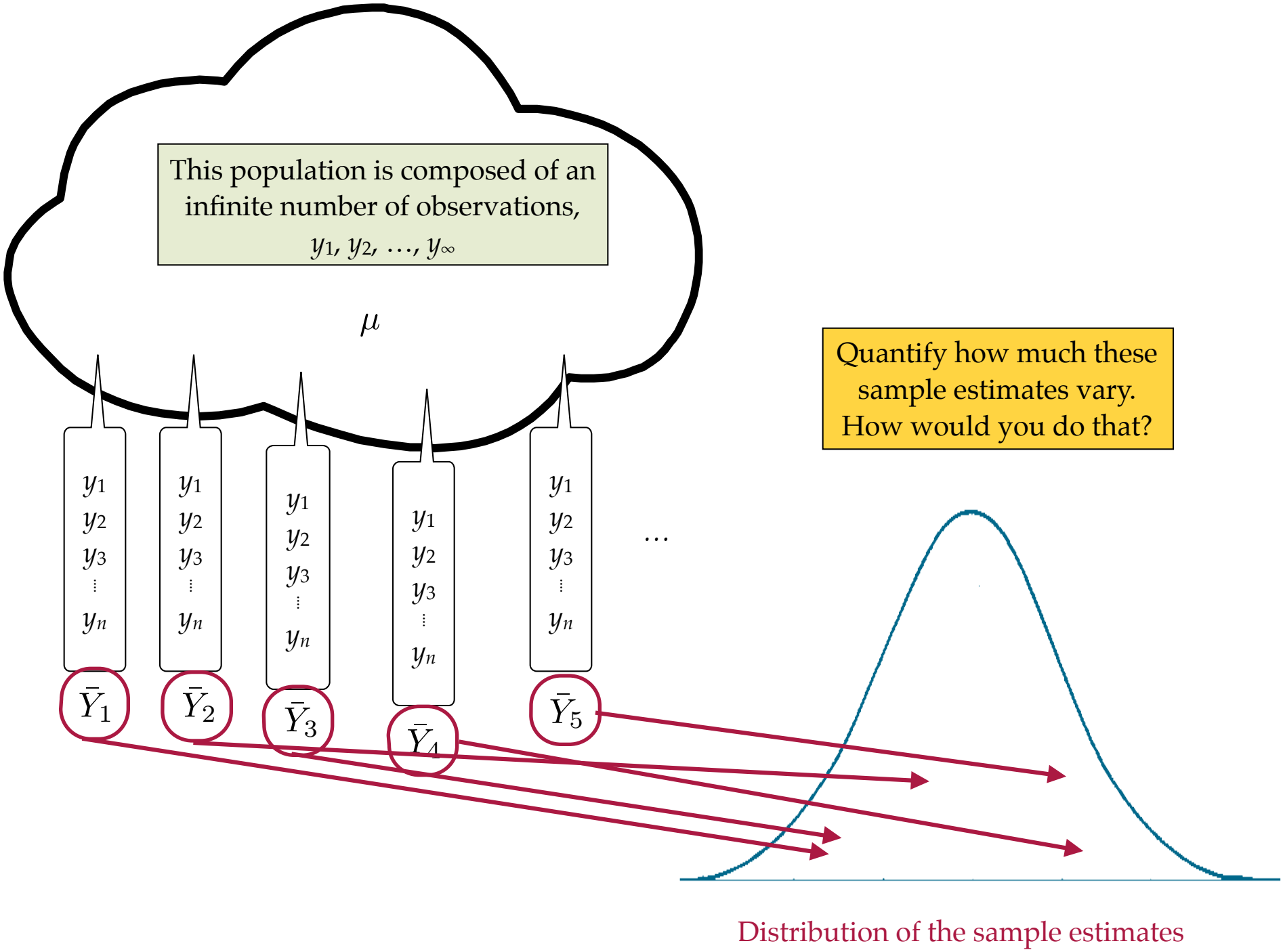
Different set of sampled data

Problem: If we had drawn a different sample of observations, we would have gotten a different estimate for the parameter!

$$\bar{Y}_1 \neq \bar{Y}_{\text{Obs}}$$

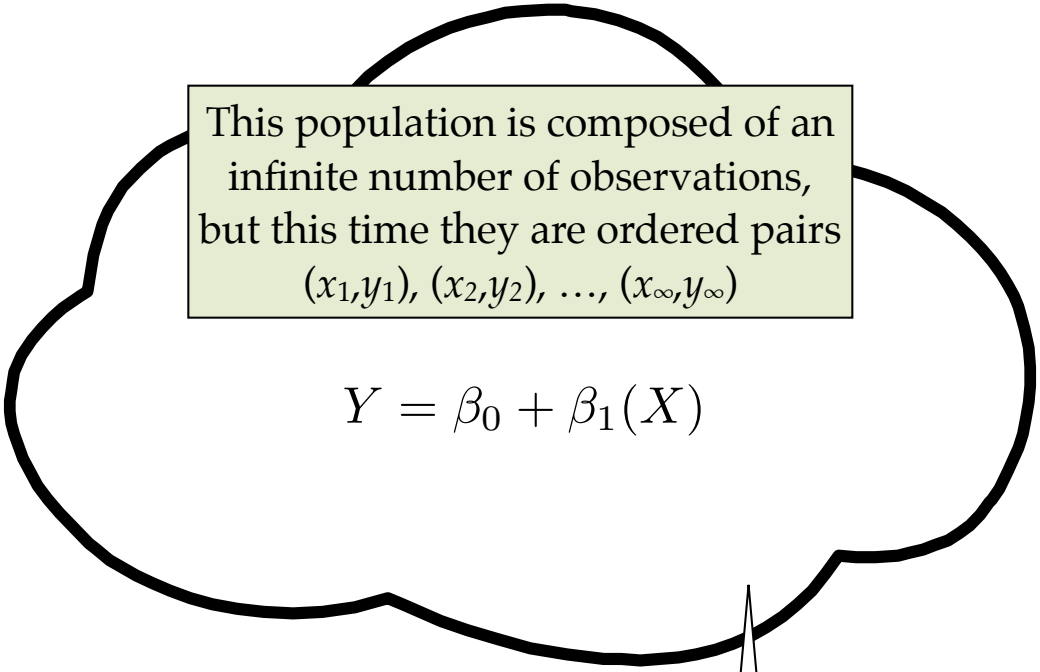
Resolution: Quantify how much these sample estimates vary across all the different samples you could possible draw.

Goal 2: Estimate the uncertainty in our estimate


This population is composed of an infinite number of observations,
$y_1, y_2, \ldots, y_\infty$
μ
y_1 y_2 y_3 ⋮ y_n
$\bar{Y}_1$
$\bar{Y}_2$
$\bar{Y}_3$
$\bar{Y}_4$
$\bar{Y}_5$
…
Quantify how much these sample estimates vary. How would you do that?
Distribution of the sample estimates

Estimation and Quantification of Uncertainty in Regression

Same Idea, Now with Regression



(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
⋮
(x_n, y_n)

$\hat{\beta}_{0_{\text{Obs}}}$

$\hat{\beta}_{1_{\text{Obs}}}$

Observed (sampled) data

Goal 1: Estimate the parameters for a given population

$$\hat{\beta}_{0_{\text{Obs}}} \approx \beta_0$$

$$\hat{\beta}_{1_{\text{Obs}}} \approx \beta_1$$

Same Problem: If we had drawn a different sample of observations, we would have gotten a different estimates for the parameters!

Same Resolution: Quantify how much these sample estimates vary across all the different samples you could possible draw.

This population is composed of an infinite number of observations, but this time they are ordered pairs (x_1,y_1), (x_2,y_2), …, (x_∞,y_∞)

$$Y = \beta_0 + \beta_1(X)$$

The distribution of the sample intercepts is normally distributed with some mean and some standard deviation.

$$\mu(\hat{\beta}_0) \qquad \sigma(\hat{\beta}_0)$$

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
⋮
(x_n, y_n)

$\hat{\beta}_{01}$

$\hat{\beta}_{11}$

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
⋮
(x_n, y_n)

$\hat{\beta}_{02}$

$\hat{\beta}_{12}$

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
⋮
(x_n, y_n)

$\hat{\beta}_{03}$

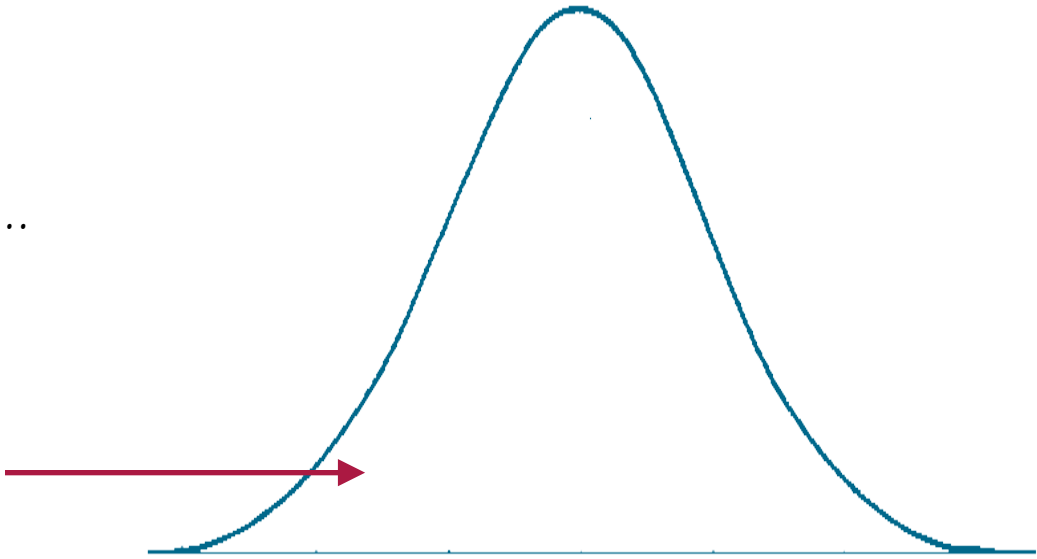
$\hat{\beta}_{13}$

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
⋮
(x_n, y_n)

$\hat{\beta}_{04}$

$\hat{\beta}_{14}$

…



Distribution of the sample intercepts

Observed (sampled) data

This population is composed of an infinite number of observations, but this time they are ordered pairs (x_1,y_1), (x_2,y_2), …, (x_∞,y_∞)

$$Y = \beta_0 + \beta_1(X)$$

The distribution of the sample slopes is normally distributed with some mean and some standard deviation.

$\mu(\hat{\beta}_1)$ $\sigma(\hat{\beta}_1)$

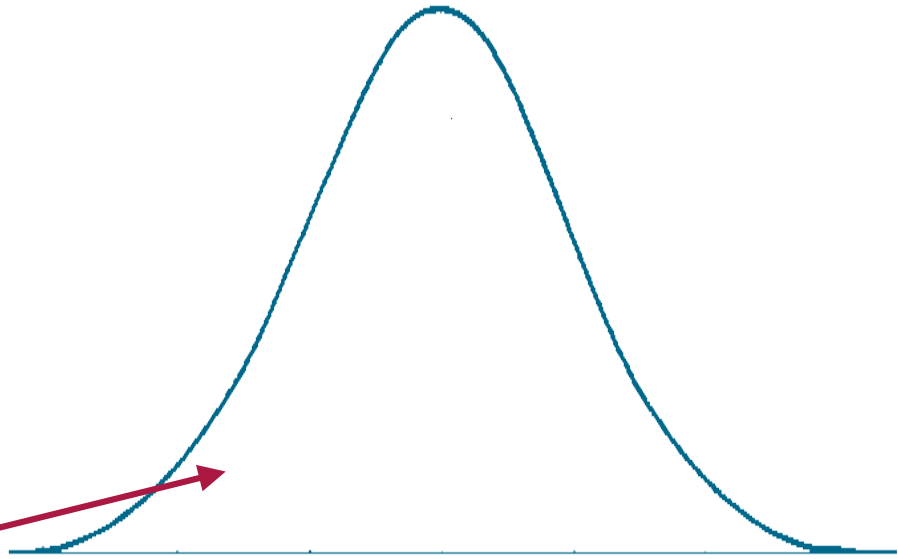
(x_1, y_1) (x_2, y_2) (x_3, y_3) ⋮ (x_n, y_n) → $\hat{\beta}_{01}$ → $\hat{\beta}_{11}$

(x_1, y_1) (x_2, y_2) (x_3, y_3) ⋮ (x_n, y_n) → $\hat{\beta}_{02}$ → $\hat{\beta}_{12}$

(x_1, y_1) (x_2, y_2) (x_3, y_3) ⋮ (x_n, y_n) → $\hat{\beta}_{03}$ → $\hat{\beta}_{13}$

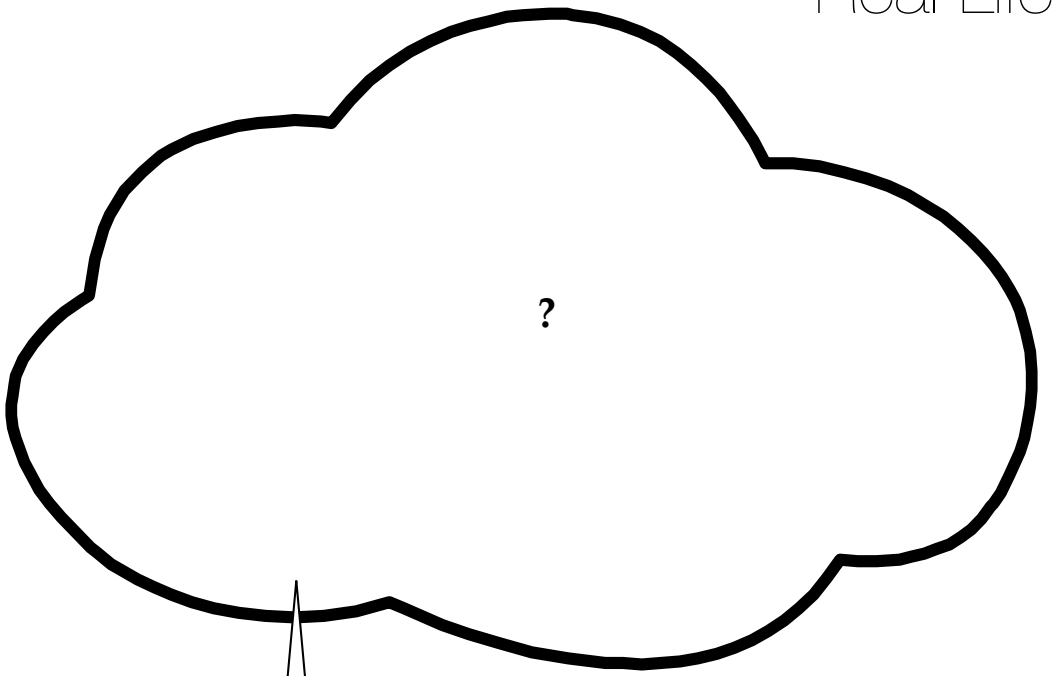
(x_1, y_1) (x_2, y_2) (x_3, y_3) ⋮ (x_n, y_n) → $\hat{\beta}_{04}$ → $\hat{\beta}_{14}$

…



Distribution of the sample slopes

Real Life



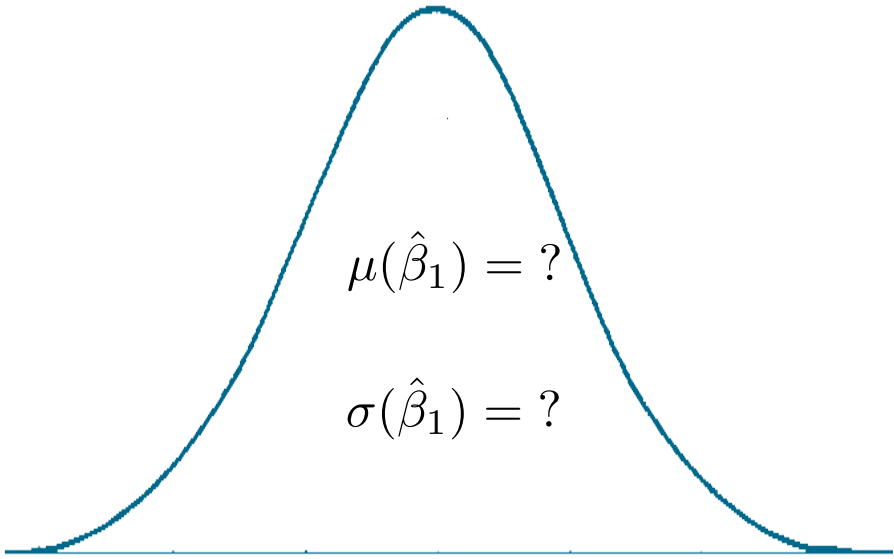
(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
$\vdots$
(x_n, y_n)

Observed (sampled) data

$\hat{\beta}_{0_{\text{Obs}}}$

$\hat{\beta}_{1_{\text{Obs}}}$

We only have a single sample… we cannot resample from the population many times.



Distribution of the sample slopes

We use the sample data to estimate both the parameter and the uncertainty.

Fit the Regression Model and Examine the Output

```
# You will need the arm package loaded to use the display() function
> display(lm.a)

lm(formula = achievement ~ homework, data = math)
            coef.est coef.se
(Intercept) 47.03     1.69
homework     1.99     0.60
---
n = 100, k = 2
residual sd = 10.75, R-Squared = 0.10
```

$$\hat{\beta}_0 = 47.03 \qquad \hat{\beta}_1 = 1.99$$

$$\hat{\sigma}(\hat{\beta}_0) = 1.69 \qquad \hat{\sigma}(\hat{\beta}_1) = 0.60$$

Slope: Using the data we have, our estimate for the population slope is 1.99. We recognize this is based on incomplete information (a sample), so we also attempt to quantify how uncertain about this estimate that we are. Based on these data, the estimate of the standard error for the slope is 0.60.

Confidence Intervals

Using the Estimate and the SE Together

$$\hat{\beta}_1 = 1.99$$

$$\hat{\sigma}(\hat{\beta}_1) = 0.60$$

Rather than present these as two separate measures, sometimes we combine them.

Estimate ± Uncertainty

$$1.99 \pm 0.60$$

$$[1.39,\ 2.59]$$

This interval combines the information in our estimate together with the amount of uncertainty. It gives us a range of candidates for the population slope.

In practice, we typically double the amount of uncertainty

$$\hat{\beta}_1 = 1.99$$

$$\hat{\sigma}(\hat{\beta}_1) = 0.60$$

$$1.99 \pm 2(0.60)$$

$$1.99 \pm 1.20$$

$$[0.79,\ 3.19]$$

The range of candidates for the population slope is now larger (interval is wider). This expresses **more uncertainty**. But, it also makes us feel **more confident** that the actual population slope (which we don't know) is one of the candidates.

Confidence Interval

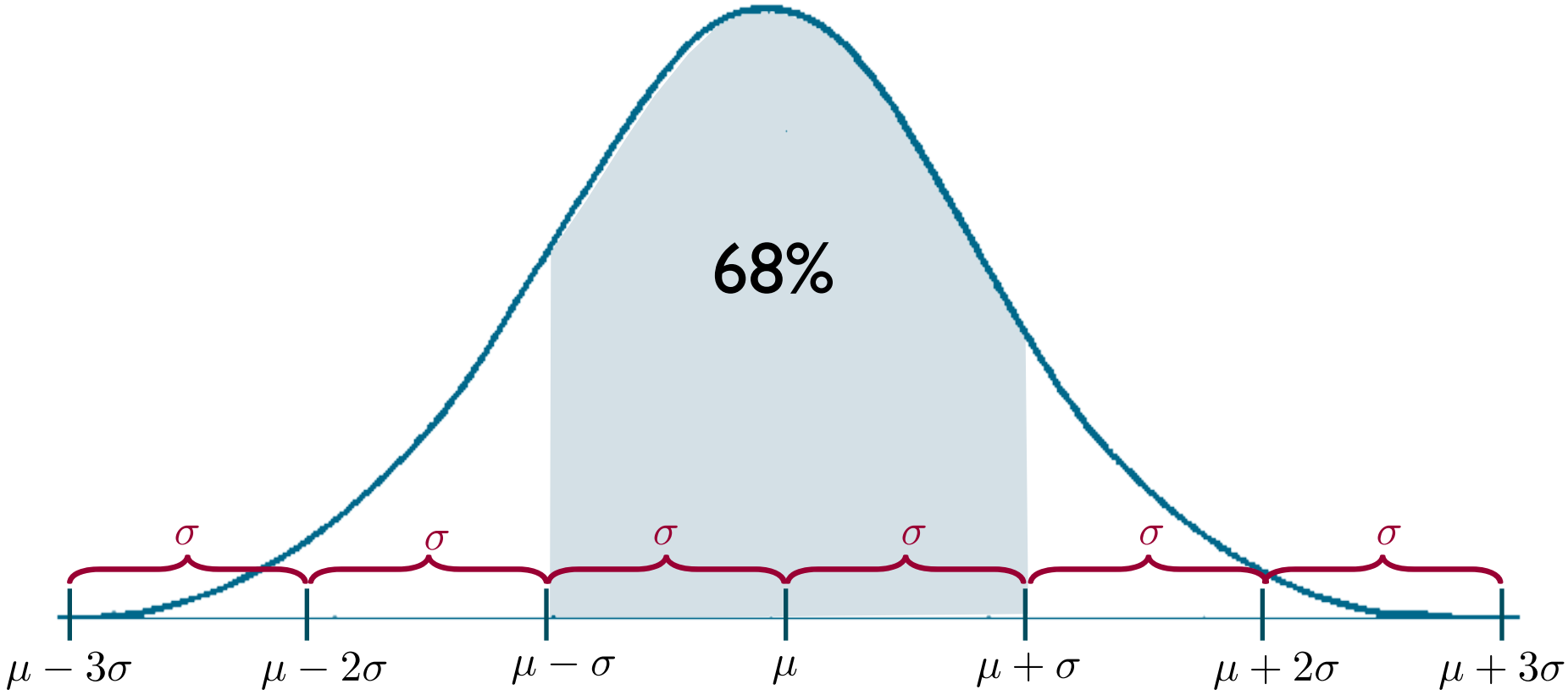
```
> confint(lm.a)

                2.5 %    97.5 %
(Intercept) 43.6698259 50.393364
homework     0.8089788  3.171389
```

The `confint()` function gives us the endpoints for the interval straightaway.

Practical Interpretation: Based on our sample data, we believe the value of the population slope is probably between 0.81 and 3.17.

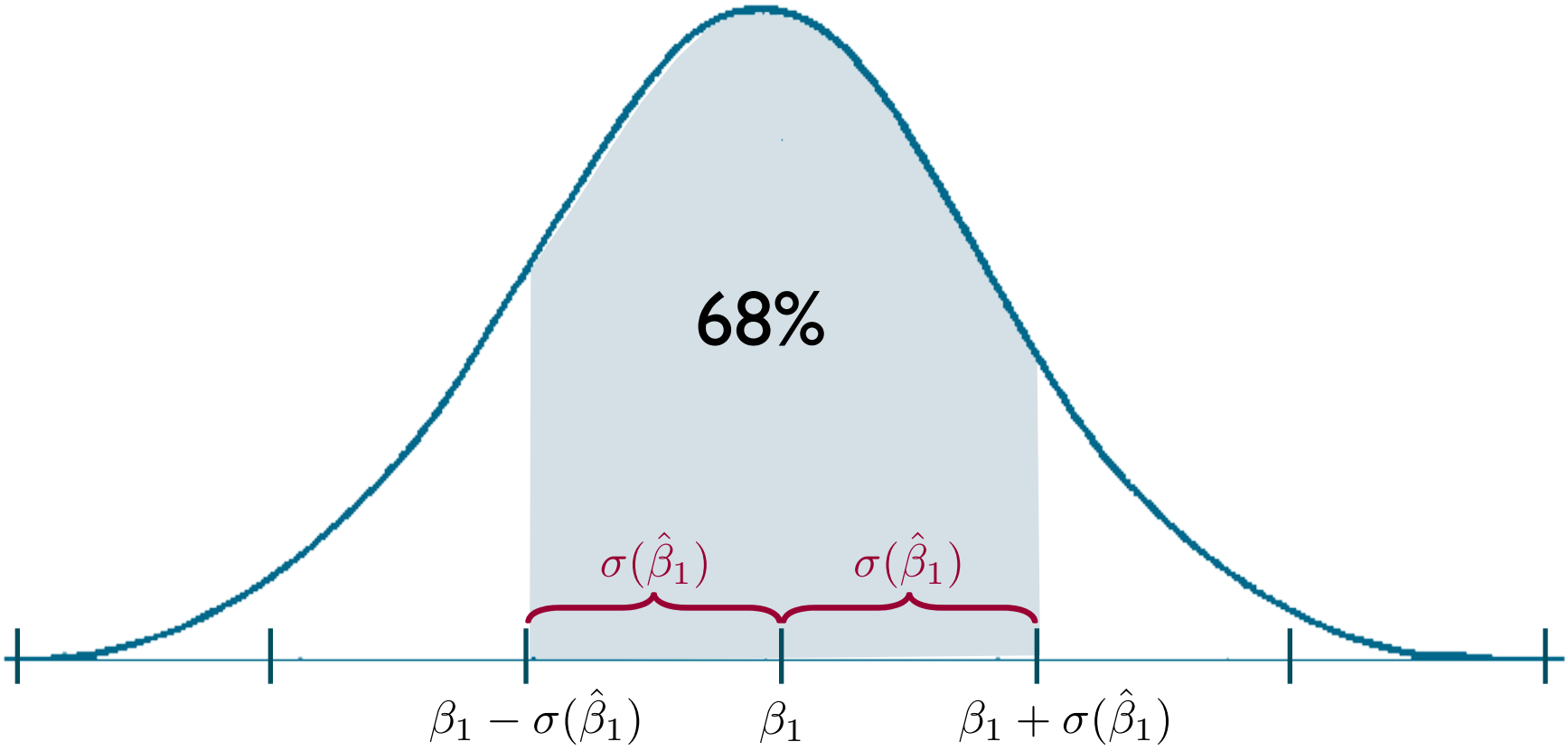
Formal Interpretation: With 95% confidence, the value of the population slope is between 0.81 and 3.17.

Some Normal Theory (A Reminder)



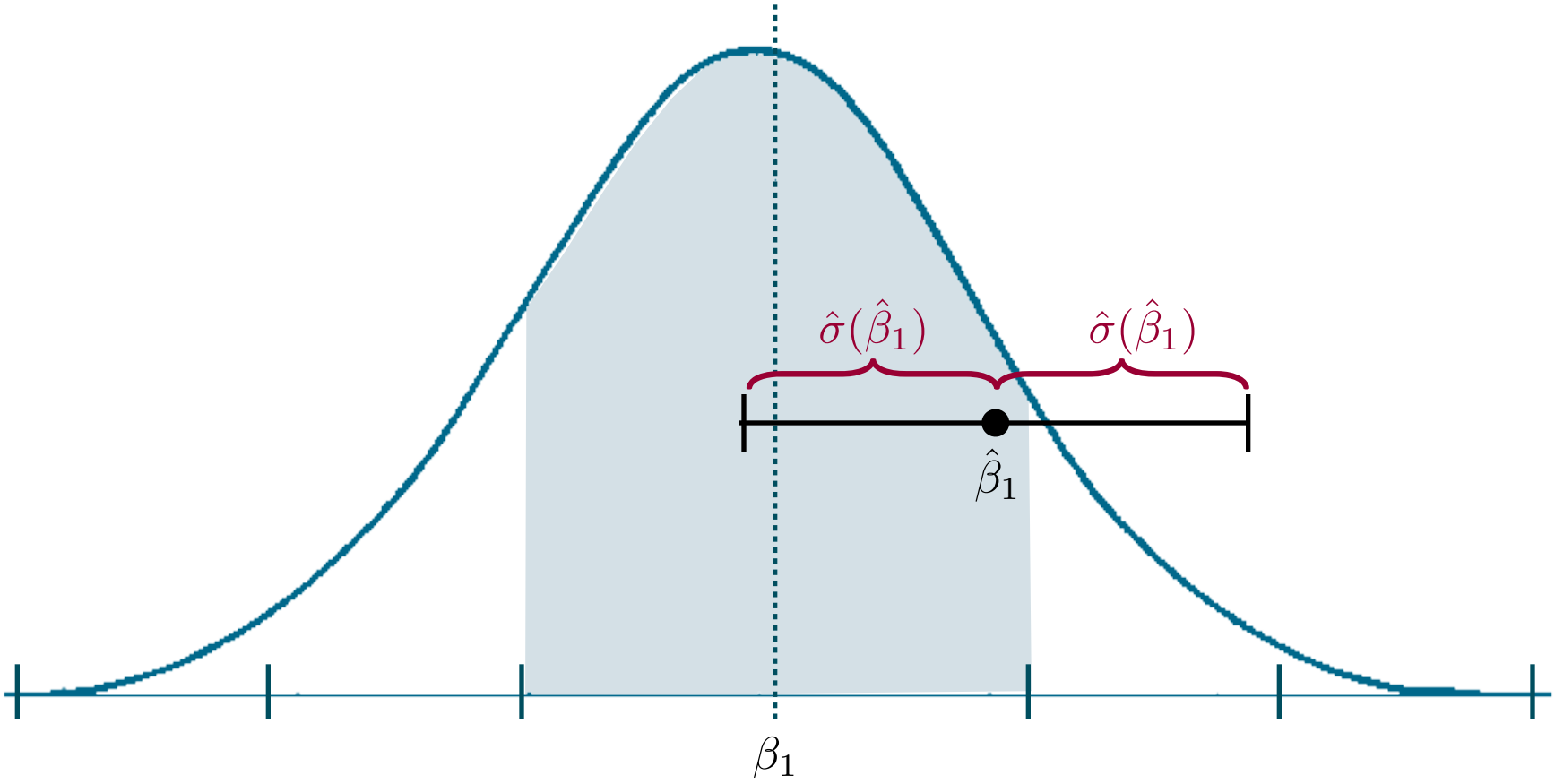
68% of the observations in a normal distribution are within one SD from the mean.

Our Distribution of Slope Estimates was also Normal…

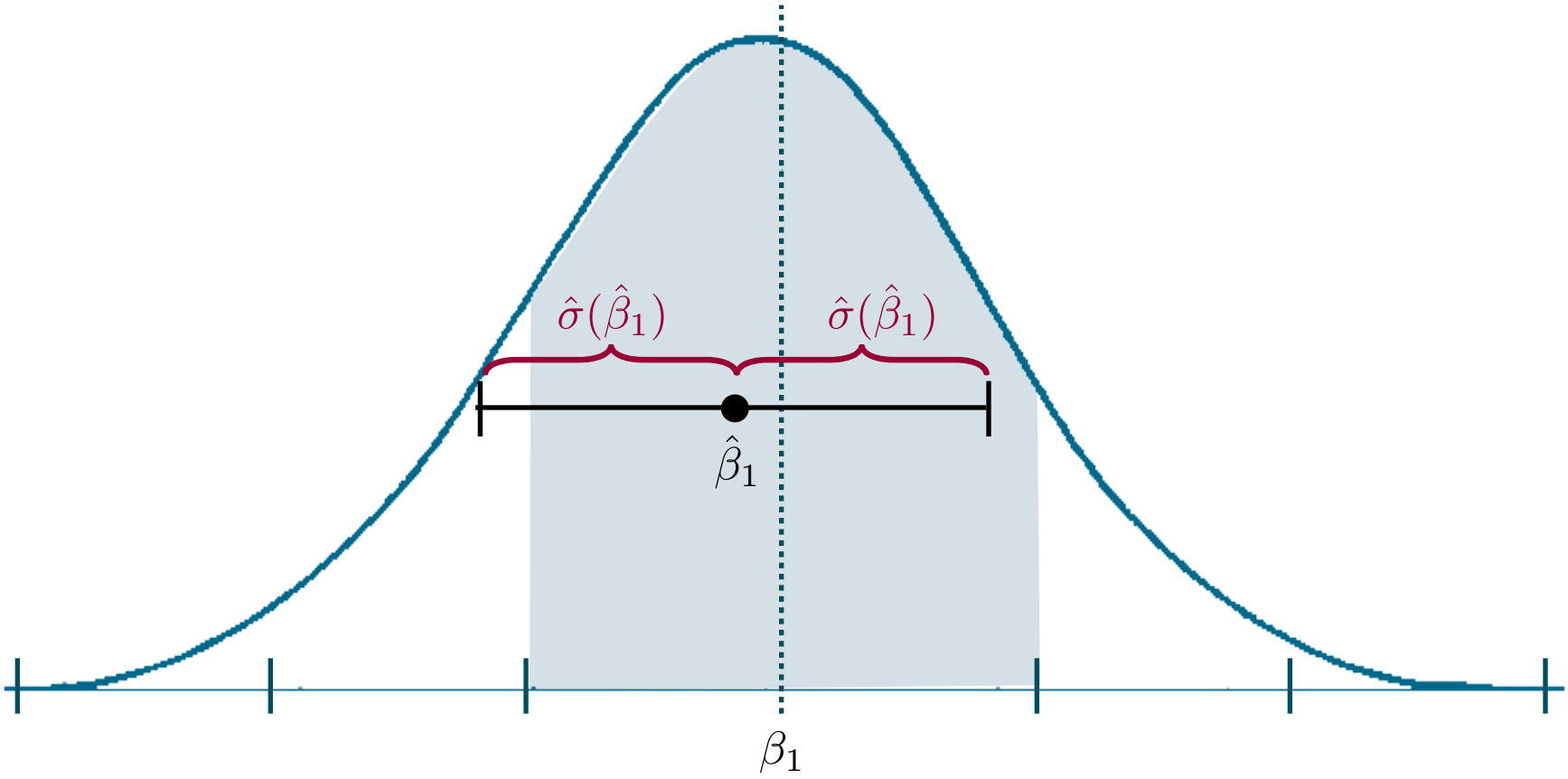


68% of all the slope estimates are within one SE from the population slope.

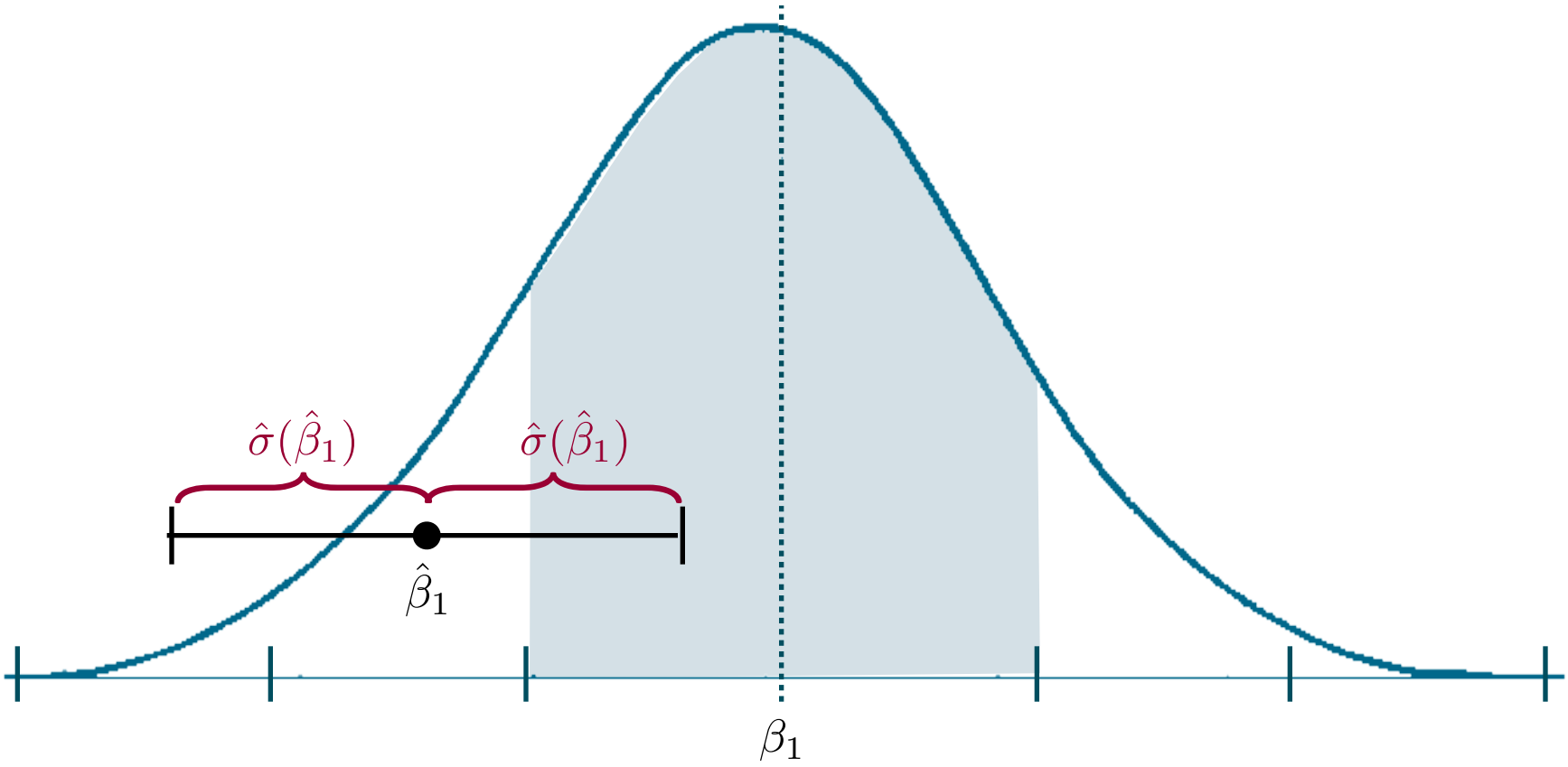
Which means we can work in the reverse…



The interval visually shows the range of candidate values for the population slope. In this interval it turns out that the actual population slope was indeed one of the possible candidates.

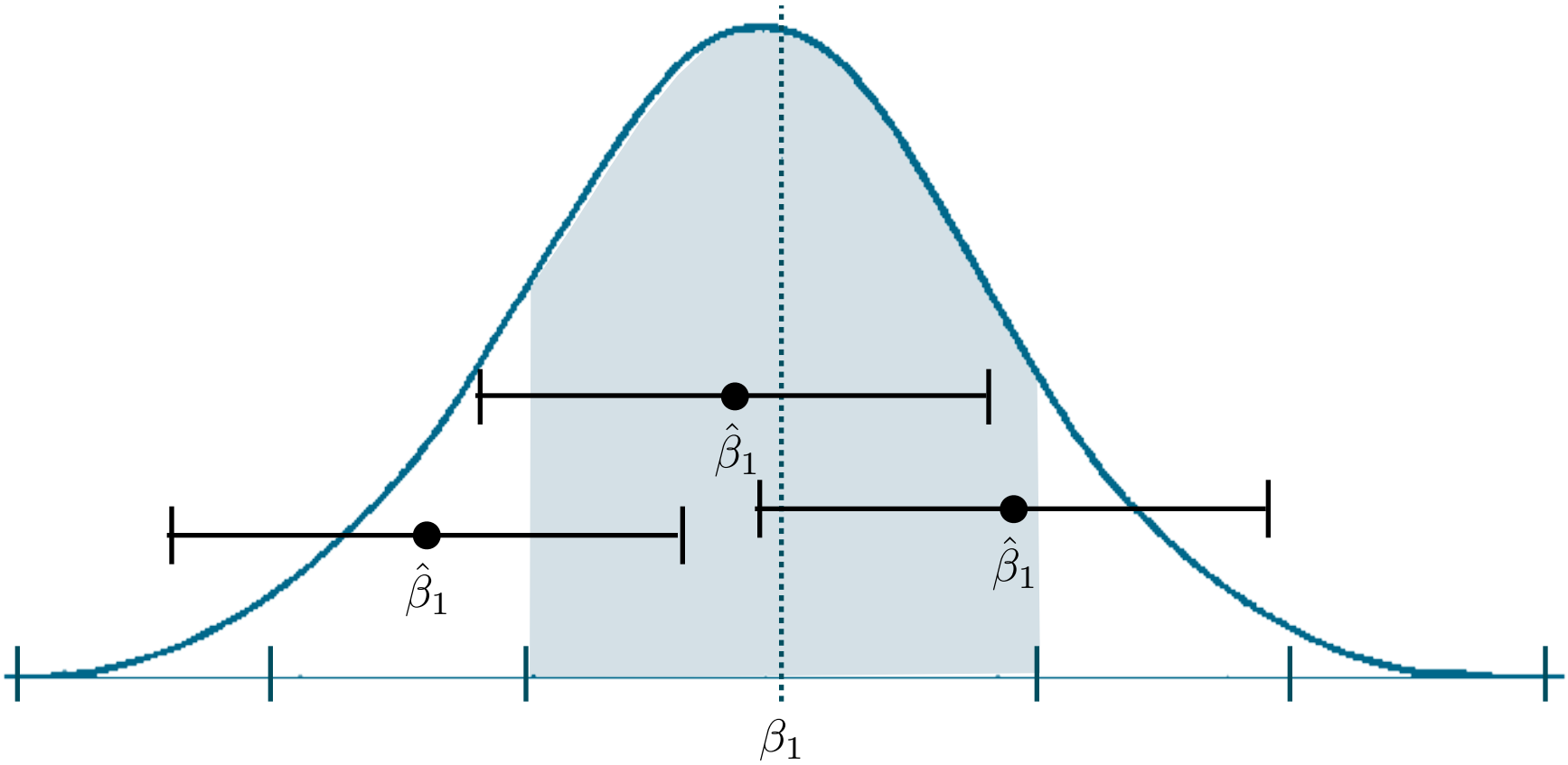


The interval visually shows the range of candidate values for the population slope. In this interval it turns out that the actual population slope was indeed one of the possible candidates.



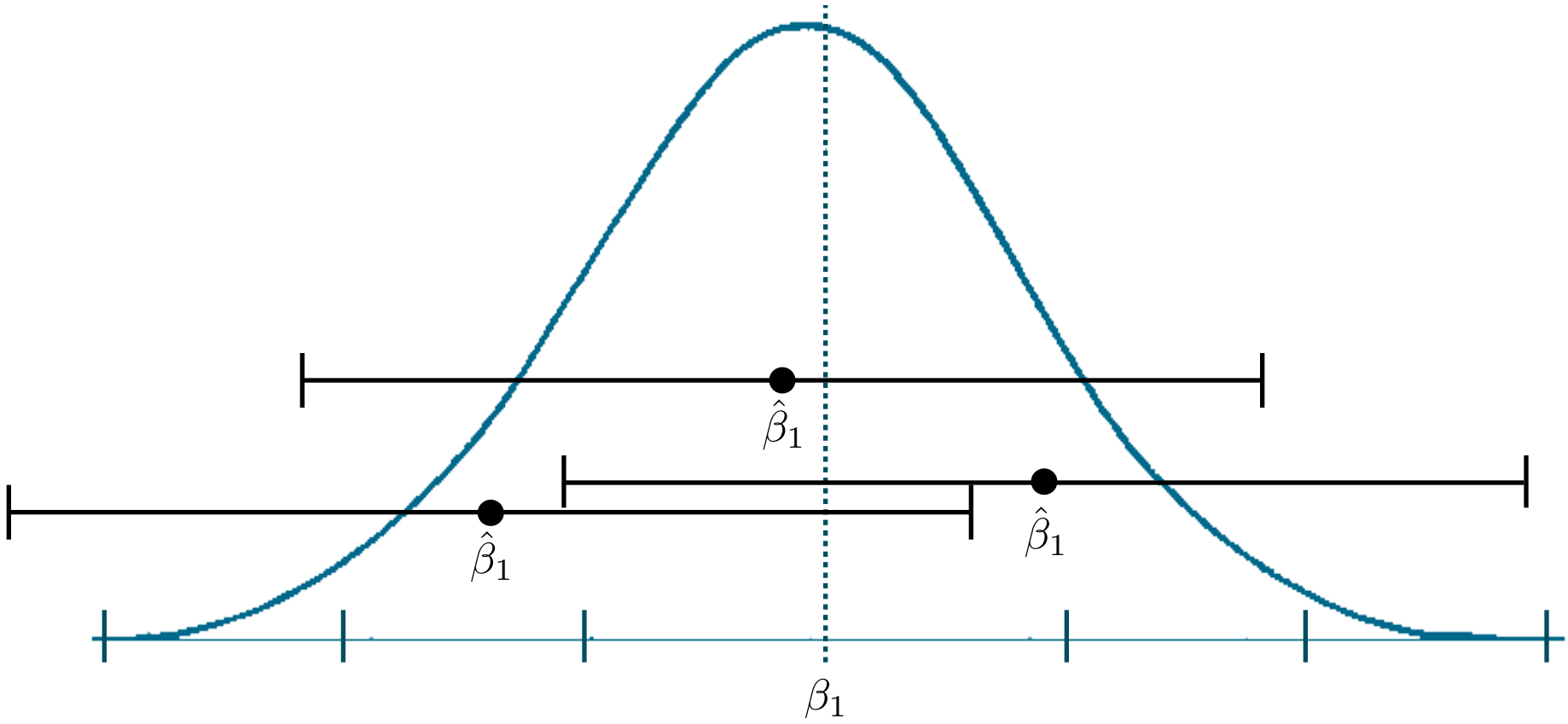
The interval visually shows the range of candidate values for the population slope. In this interval it turns out that the actual population slope was **NOT** one of the possible candidates.

Some of the slope estimates produce intervals that include the population slope in their range of candidate values. Some do not.



Consider the intervals produced for **all** of the slope estimates. What percentage would include the population slope value in their interval?

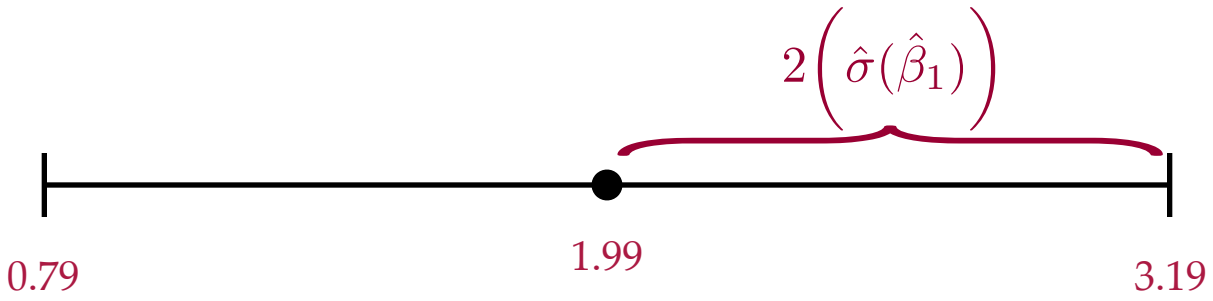
How would this change if we doubled the uncertainty to build our intervals using 2 SEs?



Actual Interpretation of Confidence Level: Of all the samples we could randomly draw from the population, 95% of them would produce an interval that includes the actual value of the population slope.

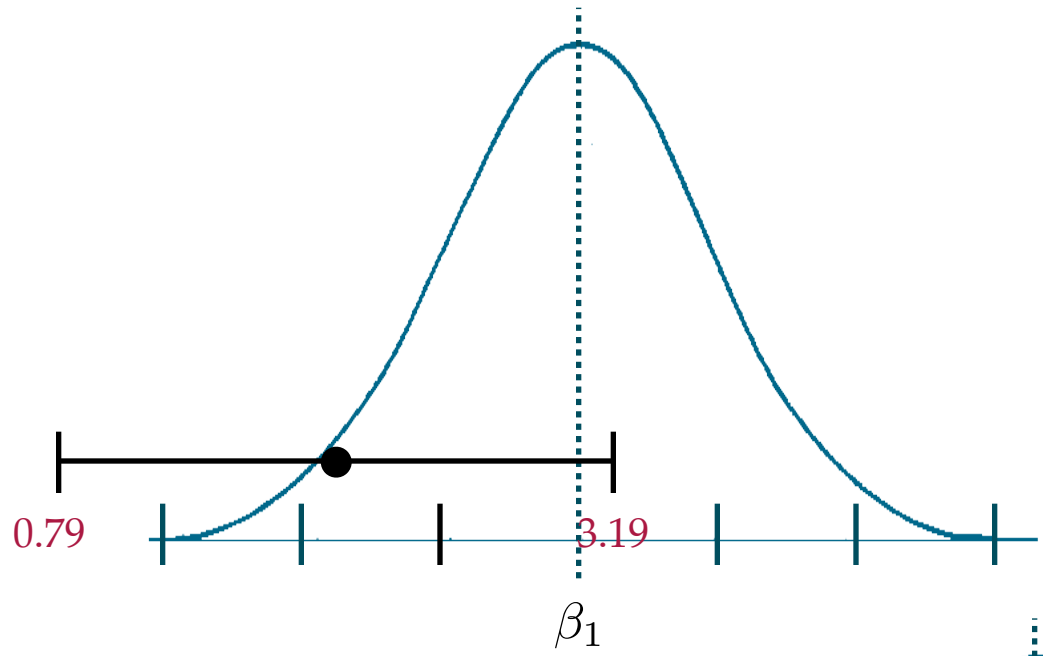
Reality

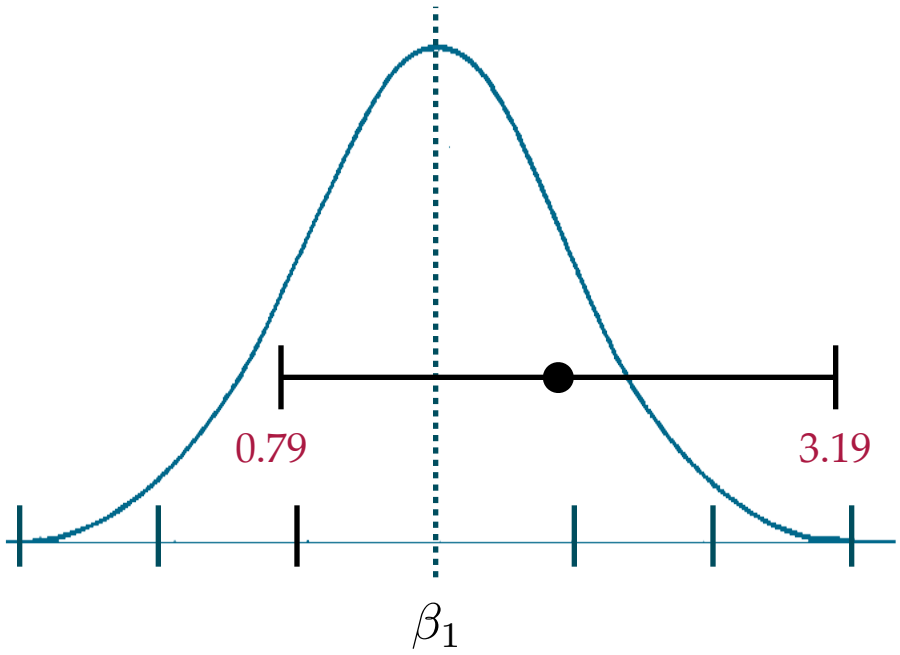
$$\hat{\beta}_1 = 1.99$$
$$\hat{\sigma}(\hat{\beta}_1) = 0.60$$

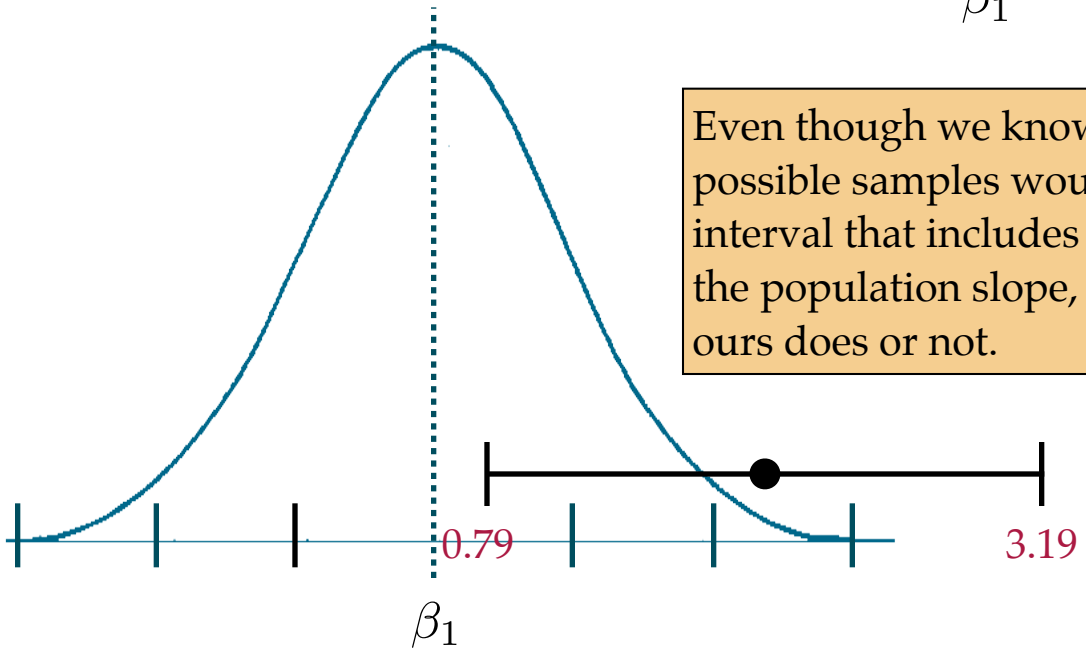


What we know is the middle and endpoints of the interval when we use our observed sample to compute the estimate and SE.

What we **do not know** is where that interval is relative to the distribution, since we do not know the value for the population slope.







Even though we know that 95% of all possible samples would produce an interval that includes the actual value of the population slope, **we have no idea** if ours does or not.

Hypothesis Testing

Testing Specific Values of the Parameter

Some research questions point to examining whether one of the regression parameters is a specific value. (e.g.,Is $\beta_1 = 0$?)

$$H_0 : \beta_1 = 0$$

We state the value we are testing in a statement called a *hypothesis*. When the value we are testing is zero, the statement is referred to as a *null hypothesis*.

It would seem logical that one could just examine the estimate of the parameter from the observed sample to answer this question…

$$\hat{\beta}_1 = 1.99$$

But, we also have to account for sampling uncertainty.

The hypothesis is a statement about the population. Here we hypothesize $\beta_1 = 0$.

$$Y = \beta_0 + (0)(X) = \beta_0$$

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
$\vdots$
(x_n, y_n)

$\hat{\beta}_{01}$

$\hat{\beta}_{11}$

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
$\vdots$
(x_n, y_n)

$\hat{\beta}_{02}$

$\hat{\beta}_{12}$

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
$\vdots$
(x_n, y_n)

$\hat{\beta}_{03}$

$\hat{\beta}_{13}$

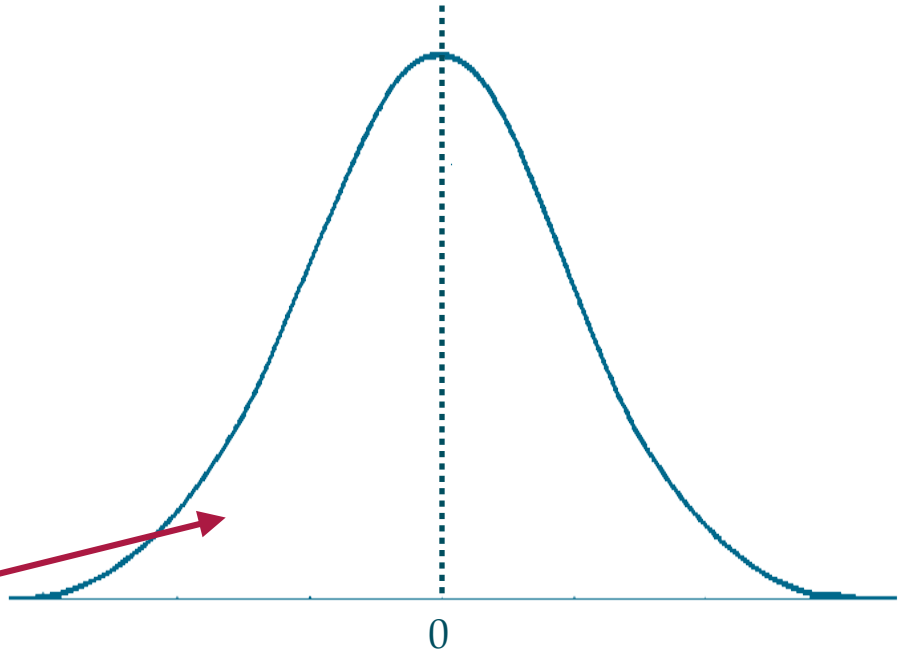
(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
$\vdots$
(x_n, y_n)

$\hat{\beta}_{04}$

$\hat{\beta}_{14}$

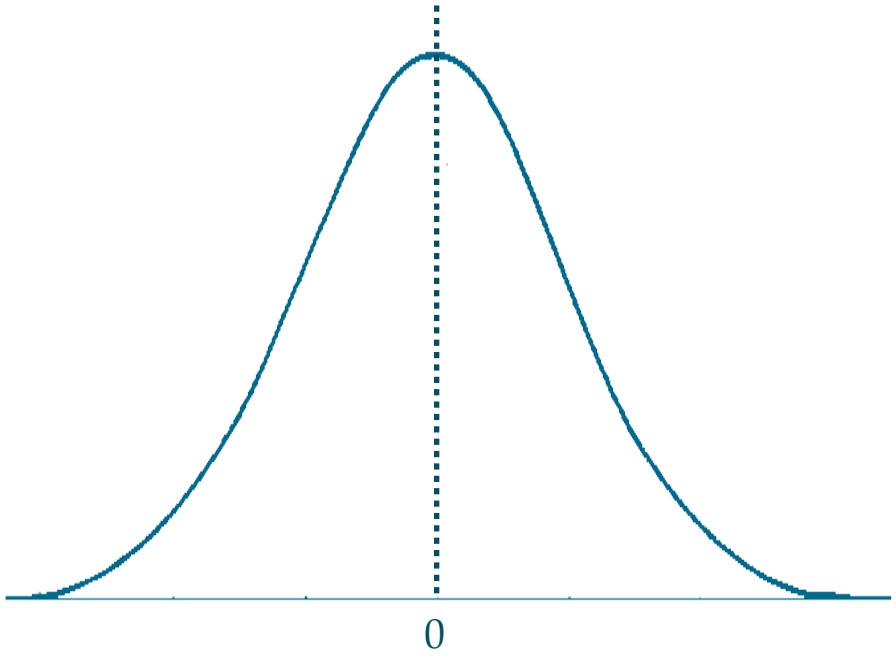
...

The distribution of the sample slopes is normally distributed with mean = $\beta_1 = 0$ and some standard deviation.



Distribution of the sample slopes

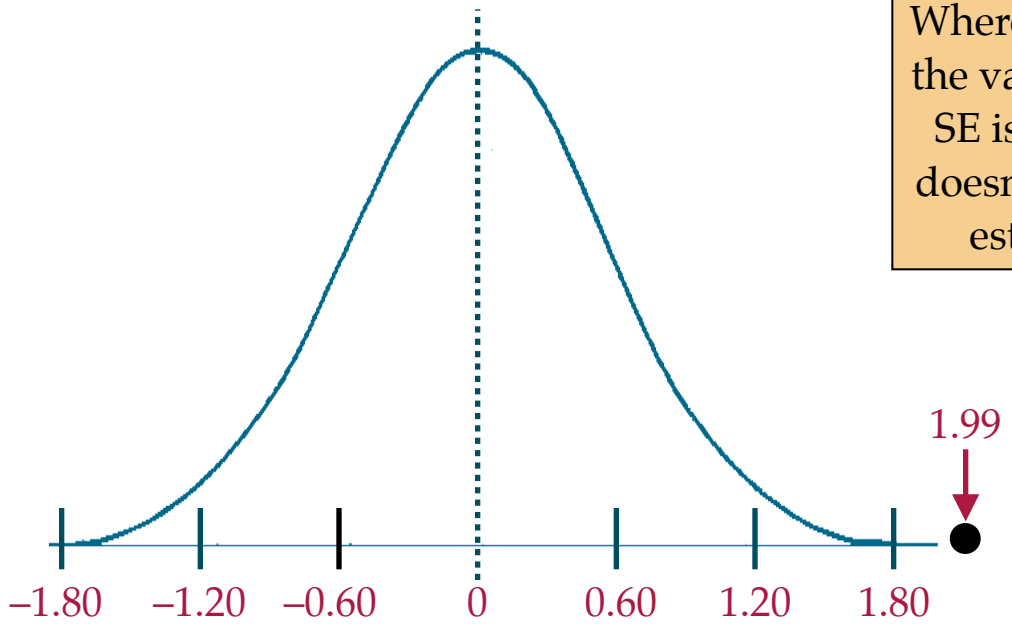
Note that even though the population has a slope of 0, we can still obtain a sample that produces an estimate for that slope that is **not** 0.



Distribution of the sample slopes

The statistical question is: If the population slope is 0, how likely is it to see an observed sample with an estimated slope of 1.99?

To answer this, essentially boils down to putting 1.99 in this distribution and then quantifying how likely that result is.



Where 1.99 falls in this distribution depends on the value of the standard error. Fortunately, the SE is primarily a function of sample size and doesn't change because of the slope value. Our estimate of 0.60 can be used here as well.

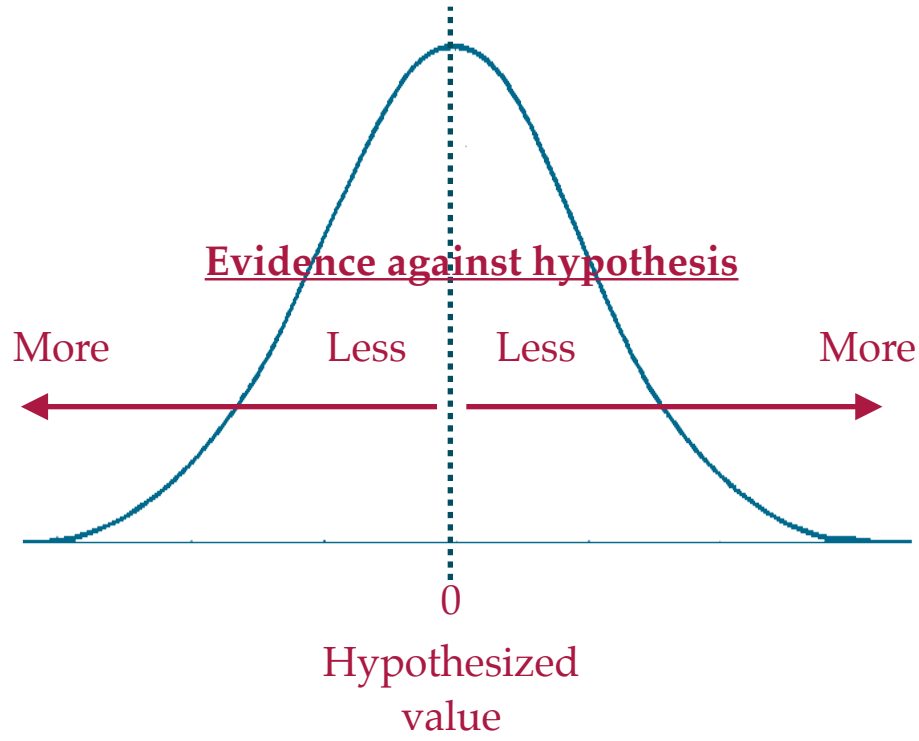
$$\frac{1.99 - 0}{0.60} = 3.31$$

The estimate of 1.99 is 3.31 SEs from the hypothesized slope value of 0.

Evidence

The observed data and estimates we get from that data are the evidence we use to judge a hypothesis.

In general, the further an estimate falls from our hypothesized value, the more evidence it provides against the hypothesis.

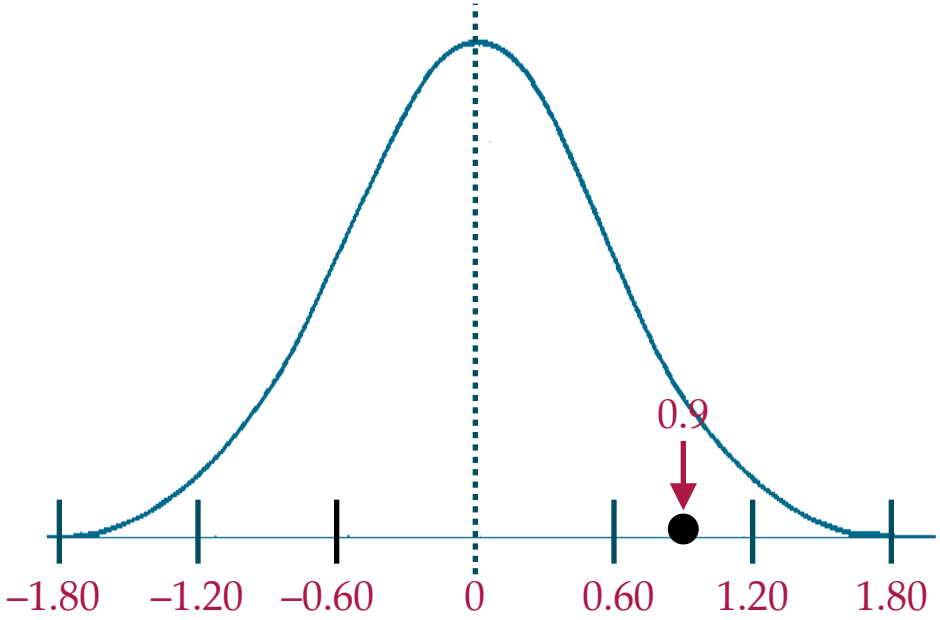


While evidence may lead us to reject a hypothesis, it can never lead us to confirm a hypothesis. A better scientific question may be, "Can I rule out this hypothesis?"

To understand how we quantify how likely this value is, let's pretend we had a slope estimate of 0.9 (instead of 1.99)

$$\frac{0.9 - 0}{0.60} = 1.5$$

The estimate of 0.9 is 1.5 SEs from the hypothesized slope value of 0.



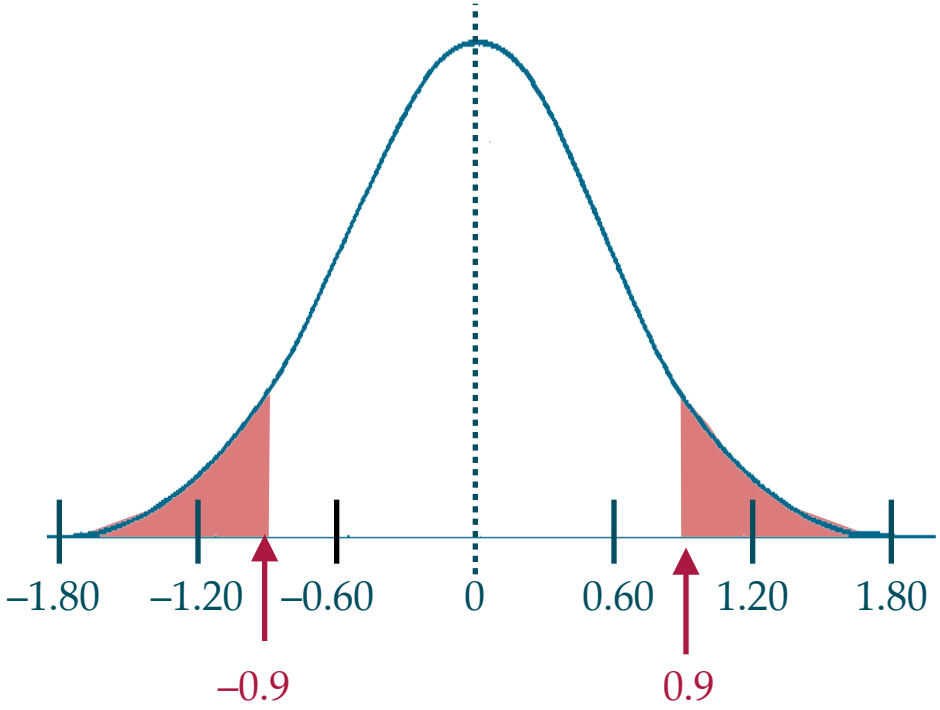
Remember, we are using 0.9 to measure the degree of evidence against the hypothesis of 0.

- If 0.9 indeed constitutes evidence against the hypothesis, values **more extreme** than 0.9 (e.g., 1.2, 3.4, 10.9, etc.) would also constitute evidence against the hypothesis.
- Similarly, since the distribution is symmetric, if we are considering evidence against the hypothesis, **regardless of which side of the distribution** the observed value is on, it constitutes that same degree of evidence against the hypothesis.

The red are constitutes all of the evidence in the distribution against the hypothesis that is at least as extreme as the observed value.

This area, called the *p*-value is 0.137 of the distribution.

Interpretation: If the hypothesis that $\beta_1 = 0$ is true, then the probability of obtaining a sample slope of *at least* 0.9 is 0.137.

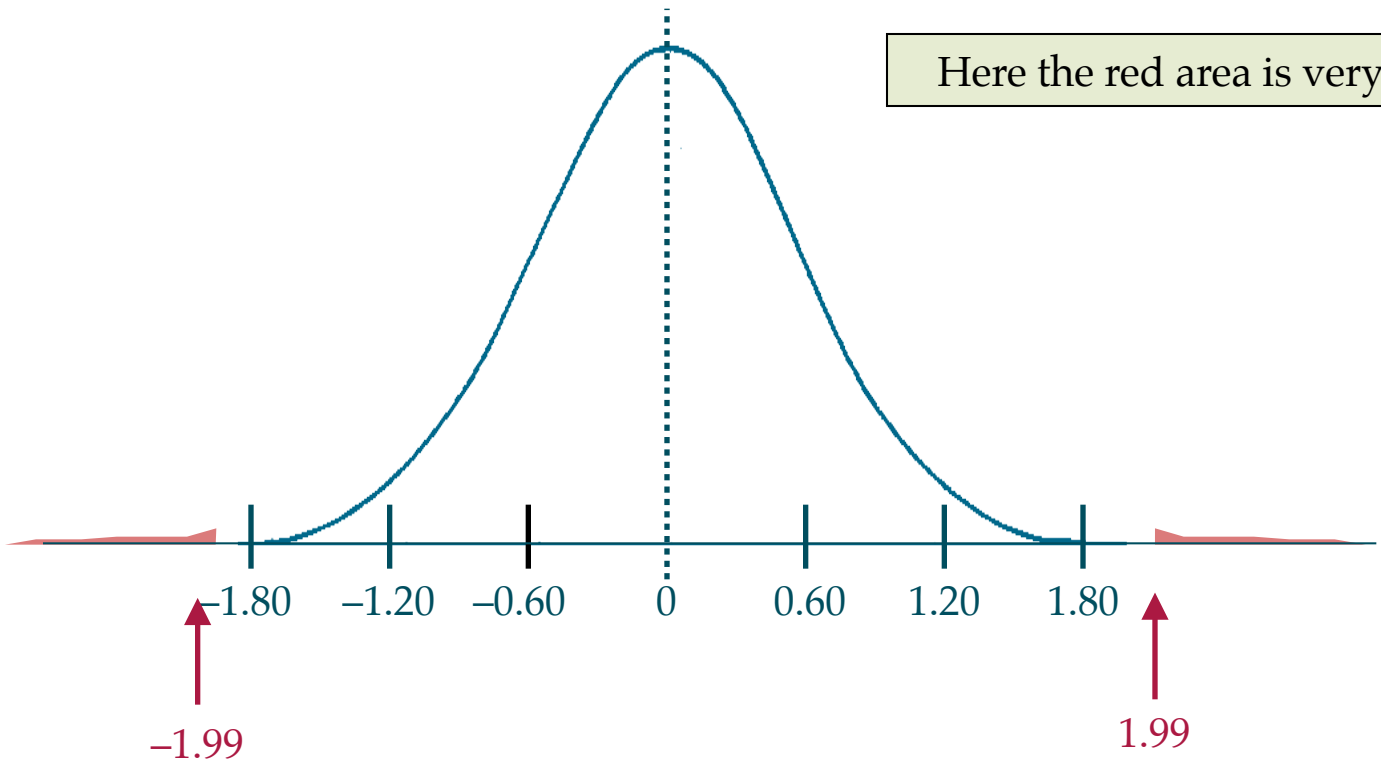


The quantification (*p*-value) we compute is based on the hypothesis being TRUE.

Recall we are asking whether we can rule out the hypothesis in question. To do this we need evidence *against* the original hypothesis.

Thus, the *p*-value is small when there is **more evidence against the hypothesis** and large when there is less evidence against the hypothesis.

Back to the Actual Data



If the hypothesis that $\beta_1 = 0$ is true, then the probability of obtaining a sample slope of at least 1.99 is 0.001.

How much evidence do we need to reject the hypothesis?

This constitutes a great deal of evidence against the hypothesis, and **in practice**, we would reject the hypothesis that $\beta_1 = 0$.

Model Summary

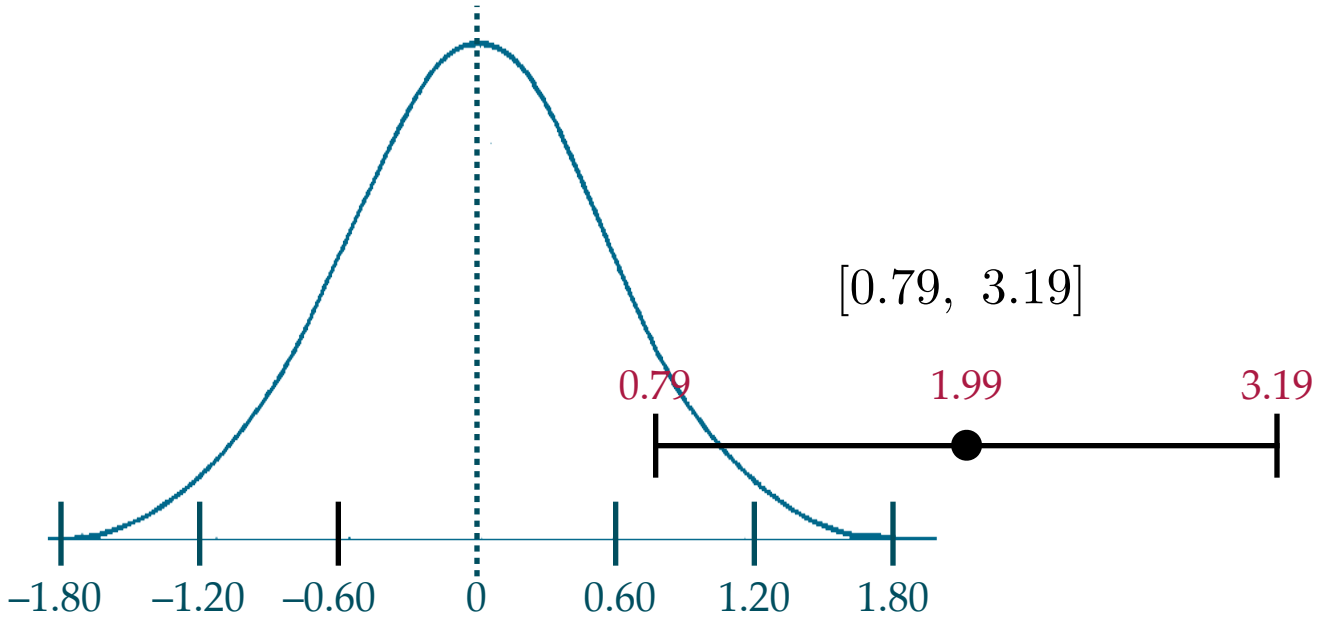
We use the `summary()` function to get coefficient estimates, SE estimates, and p-values.

```
> summary(lm.a)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.0316     1.6940  27.763  < 2e-16 ***
homework      1.9902     0.5952   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-values given in software are typically tests of the hypothesized value of 0.

Confidence Interval (Revisited)

The confidence interval gave us the range of candidate values for the population slope.



Is the hypothesized value of 0 in the range of candidate values for the population slope?

$$H_0 : \beta_1 = 0$$

Model-Level Inference

Sometimes you may want to carry out inference for the model as a whole, rather than for the individual parameters.

$$H_0 : \beta_0 = \beta_1 = 0$$

This allows for a simultaneous test of all of the parameters in the model. (It is akin to the omnibus test we carry out in ANOVA.) The model-level inference is shown in the bottom line of the `summary()` output.

```
> summary(lm.a)

F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

Here we would conclude that at least one of the parameters in the model is not zero.

Testing Variance Accounted For

Another way we can conceptualize the model-level hypothesis test is

$$H_0 : \rho^2 = 0$$

We use the same results from the model-level *F*-test. The summary results also tell us the sample estimate of rho-squared.

```
> summary(lm.a)

Multiple R-squared:  0.1024,   Adjusted R-squared:  0.09324
F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

Here the sample estimate, *R*2, suggests the model accounts for 10.2% of the variation in mathematics achievement scores, *in the sample*. Does the model account for variation in achievement scores *in the population*? Or is the sample result only due to sampling error?

Same Result for Test of Slope and Model???

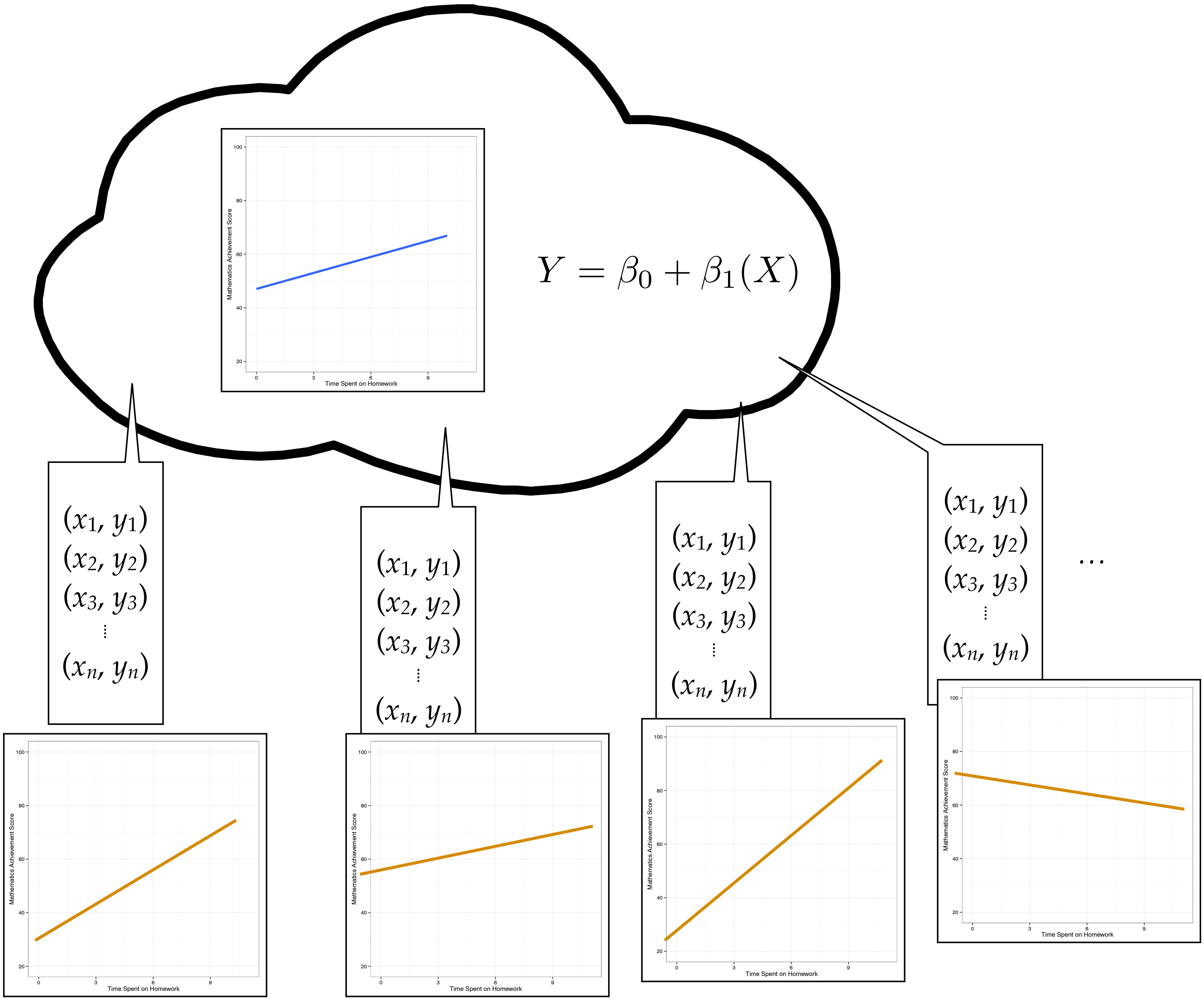
```
> summary(lm.a)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.0316     1.6940  27.763  < 2e-16 ***
homework      1.9902     0.5952   3.344  0.00117 **

Multiple R-squared:  0.1024,   Adjusted R-squared:  0.09324
F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

In simple regression models, the test for the model is exactly the same as the test for the slope.

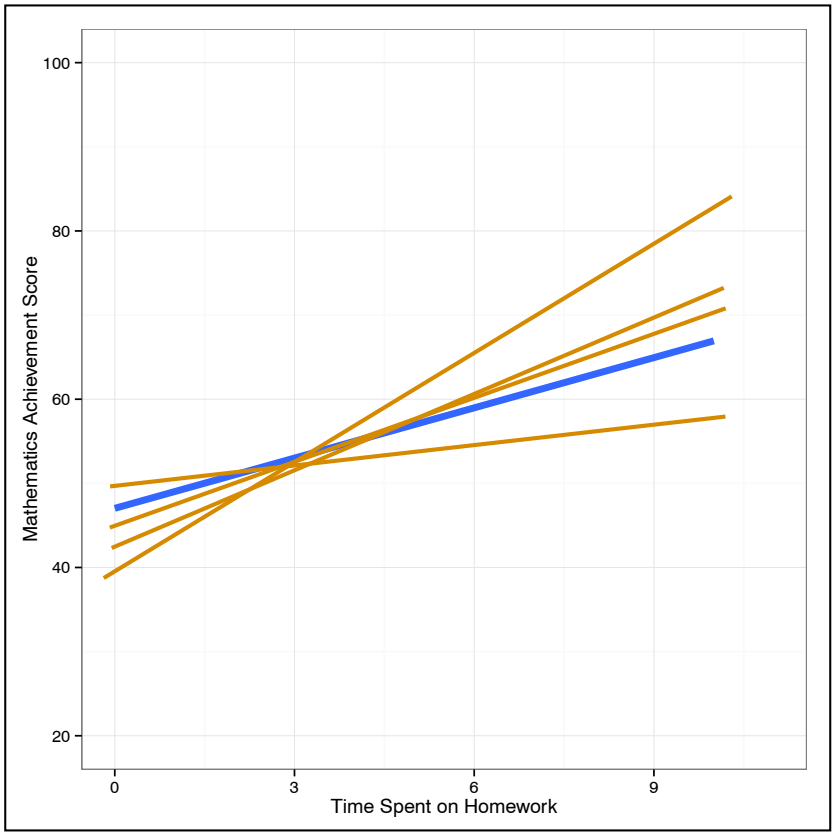
That is because the model is composed of a single predictor, so asking whether the *model accounts for variation* in achievement scores **is the same as** asking whether *differences in time spent on homework account for variation* in achievement scores.

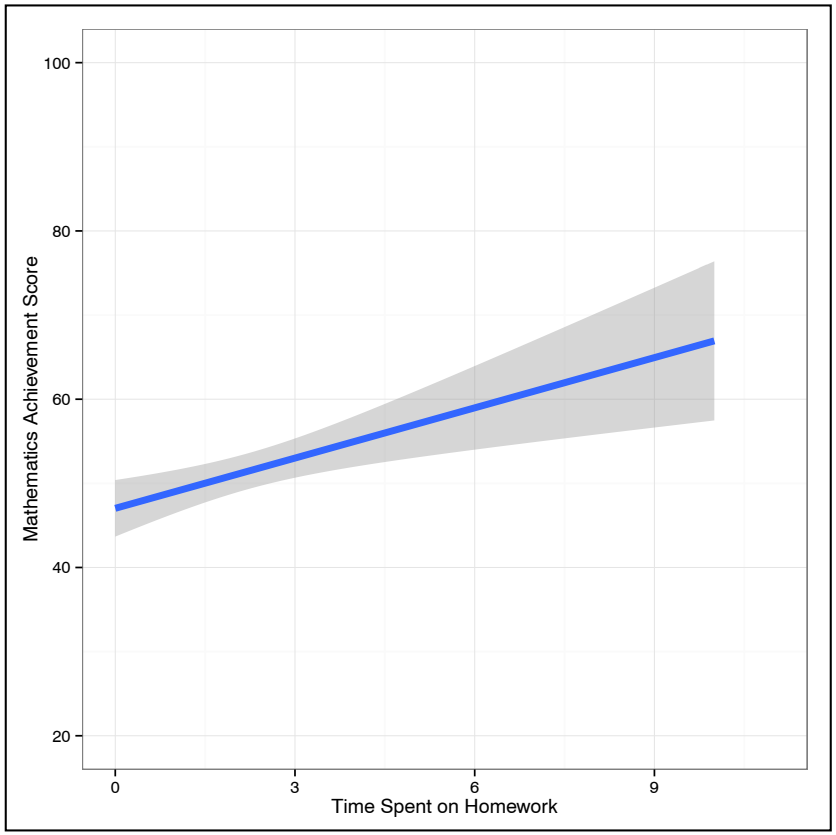
Once we have multiple predictors in the model, the model-level results and predictor-level results will not be the same.

Confidence Envelope for the Model



As we superimpose these lines on the same plot we can visualize the uncertainty in the model (i.e., in the intercept and slope).





In practice, we estimate the uncertainty from the sample data.

To plot the simple regression model and the model uncertainty, we use the `geom_smooth()` function. The `method=` argument is `"lm"` to estimate the regression model, and `se=TRUE` adds the confidence envelope. If you only want the line, use `se=FALSE`.

```
> ggplot(data = math, aes(x = homework, y = achievement)) +
    geom_smooth(method = "lm", se = TRUE) +
    xlab("Time Spent on Homework") +
    ylab("Mathematics Achievement Score") +
    theme_bw()
```