REGRESSION REVIEW

Fundamentals of

PROGRAM EVALUATION

JESSE LECY

THE ROAD MAP

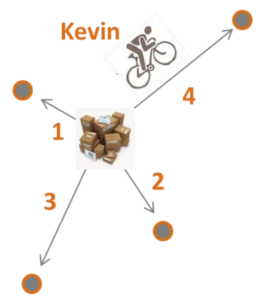
	Of the Mean:	Of the Slope:
Variance:	$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ (for x)	$\sigma_\varepsilon^2 = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}$ (using the residual)
↓		
Standard Deviation:	$\sigma_x = \sqrt{\sigma_x^2}$	$\sigma_\varepsilon = \sqrt{\sigma_\varepsilon^2}$
↓		
Standard Error:	$SE_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$	$SE_{b_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$
↓		
Confidence Interval	$\mu = \bar{x} \pm t \cdot SE_{\bar{x}}$ (of the mean)	$\beta_1 = b_1 \pm t \cdot SE_{b_1}$ (of the slope)

All of the statistical concepts that you have learned in the previous course using variance, standard errors, and confidence intervals of a estimates of the mean from a single variable apply to regression, but they have to be adapted.

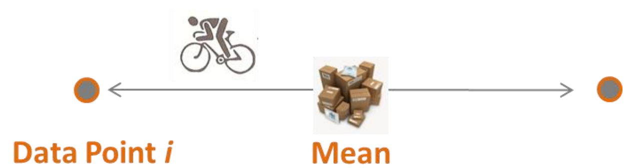
Make note that statistical concepts always need to be followed by the phrase “of the” because they are general concepts and the specific calculations are determined by the variables you are working with. The standard error around an estimated mean is different than the standard error around an estimated slope.

USEFUL METAPHORS

Variance

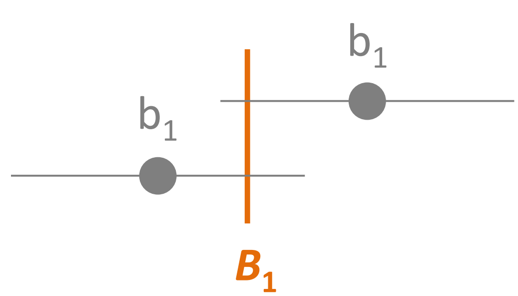


Standard Deviation



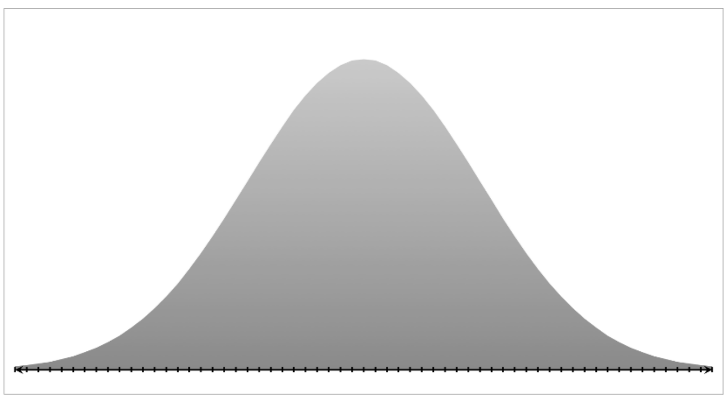
Standard Error

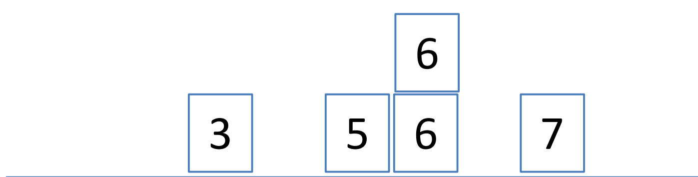
Confidence Interval



SAMPLING DISTRIBUTIONS

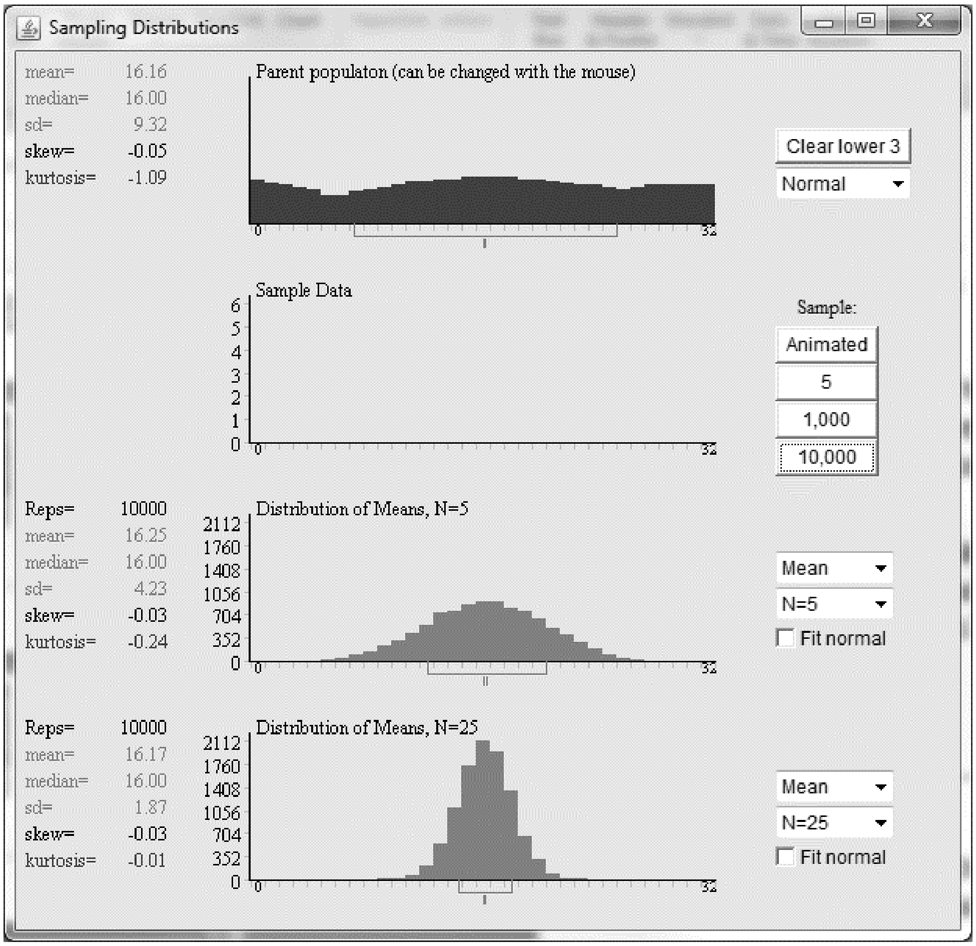
Population Statistic → $\mu = 5$





Sample Statistic → $\overline{x} = 5.4$

STANDARD ERROR OF A SAMPLE MEAN

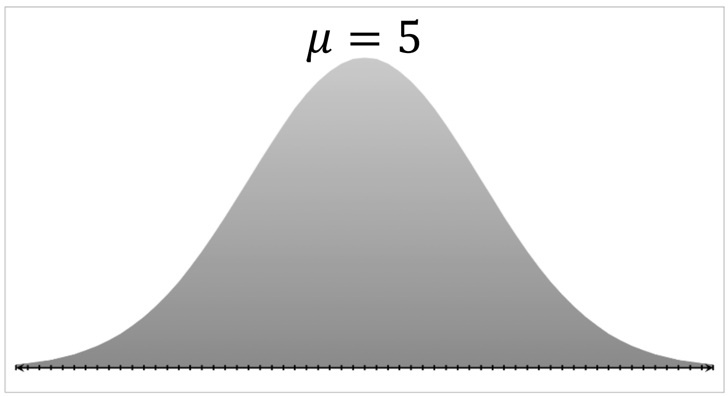


http://onlinestatbook.com/stat_sim/sampling_dist/index.html

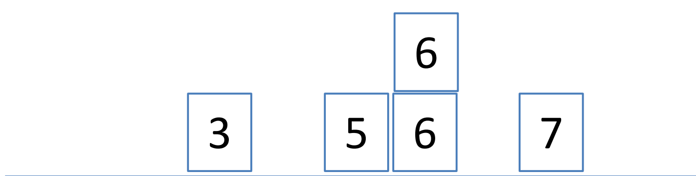
$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

STANDARD ERROR OF A SAMPLE MEAN

Population:



Sample size = 5



$$\frac{3+5+6+6+7}{5} = 5.4$$

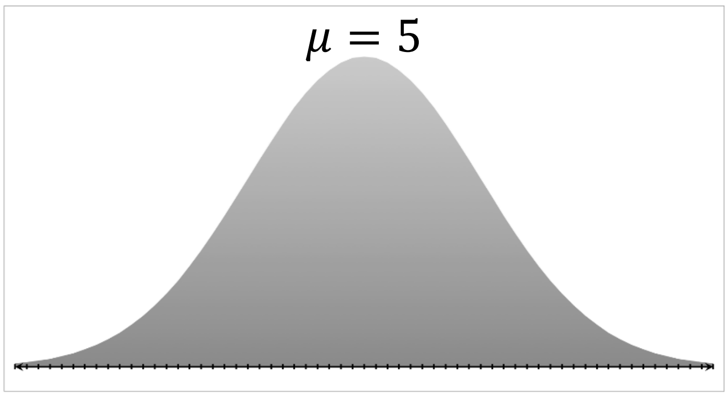
$$\mu = 5 \qquad \bar{x} = 5.4$$

How far, on average, will our best guess be from the true mean?

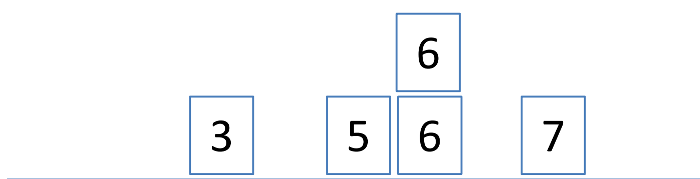
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

STANDARD ERROR OF A SAMPLE MEAN

Population:



Sample size = 5



$$\frac{3+5+6+6+7}{5} = 5.4$$

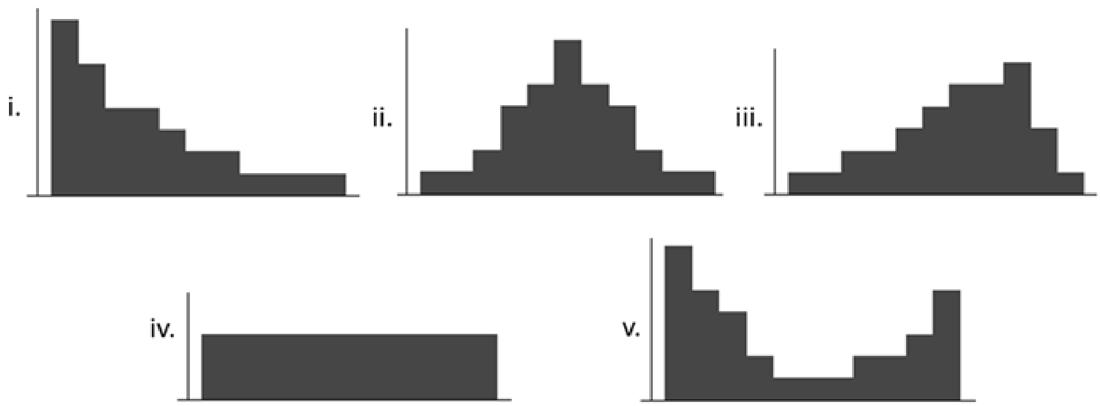
How far, on average, will our best guess be from the true mean?

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

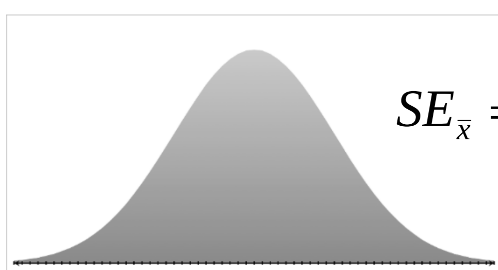
STANDARD ERROR → **"AVERAGE ERROR" (OF THE SAMPLE STAT)**

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

CENTRAL LIMIT THEOREM
ASIDE



NO MATTER WHAT THE POPULATION LOOKS LIKE



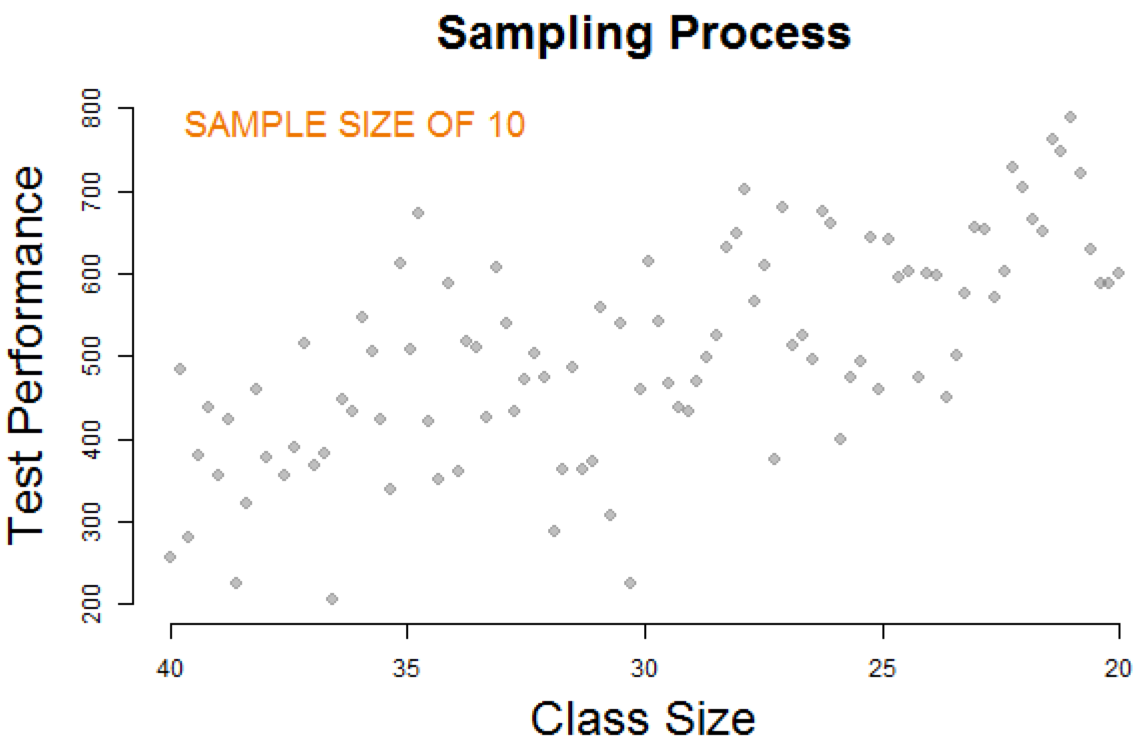
$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

THE SAMPLING DISTRIBUTION OF THE MEAN IS ALWAYS NORMAL

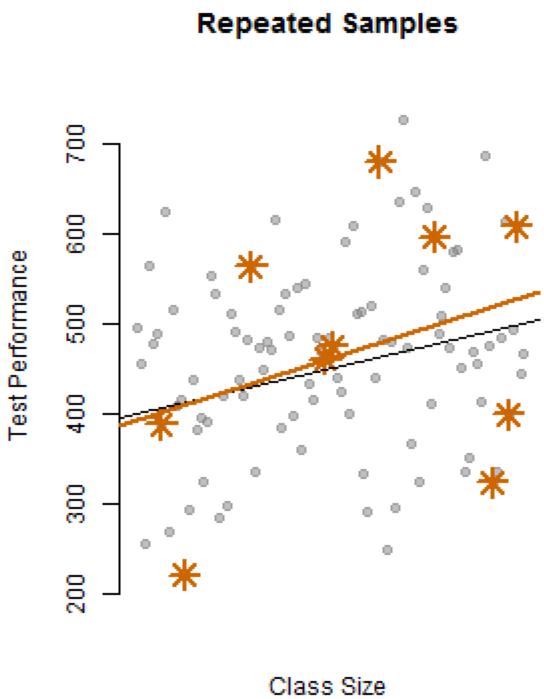
(otherwise we would not have inferential statistics)

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

STANDARD ERROR OF THE SLOPE

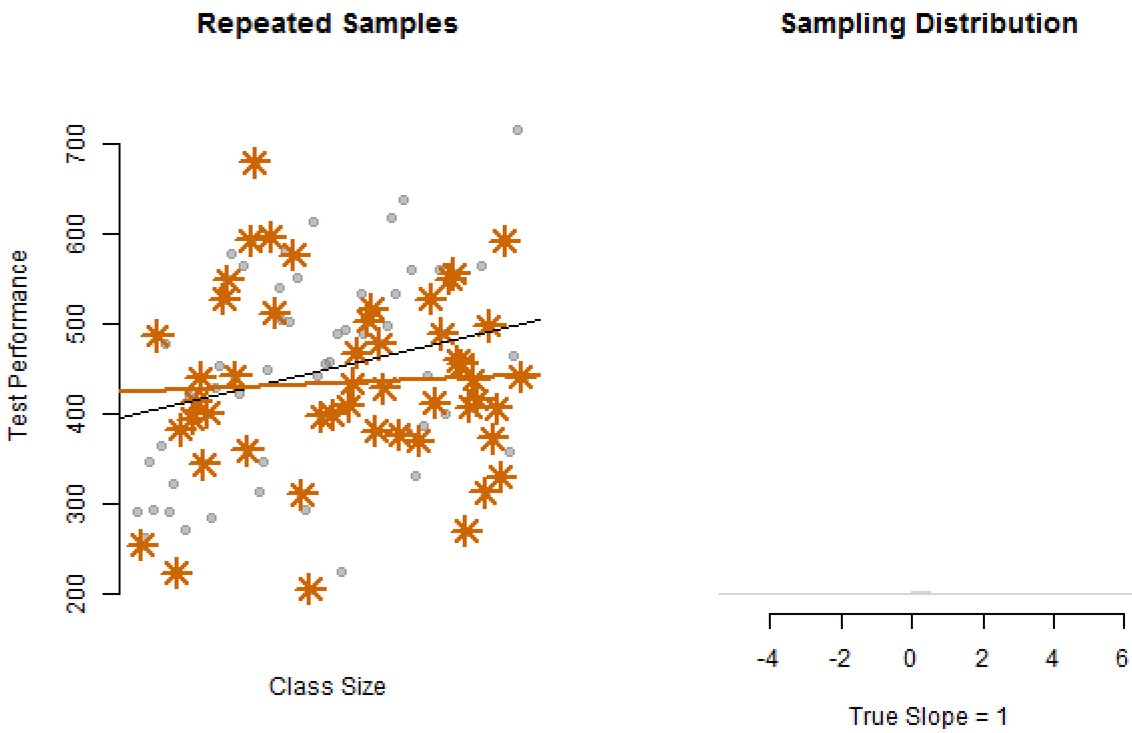


STANDARD ERROR OF THE SLOPE



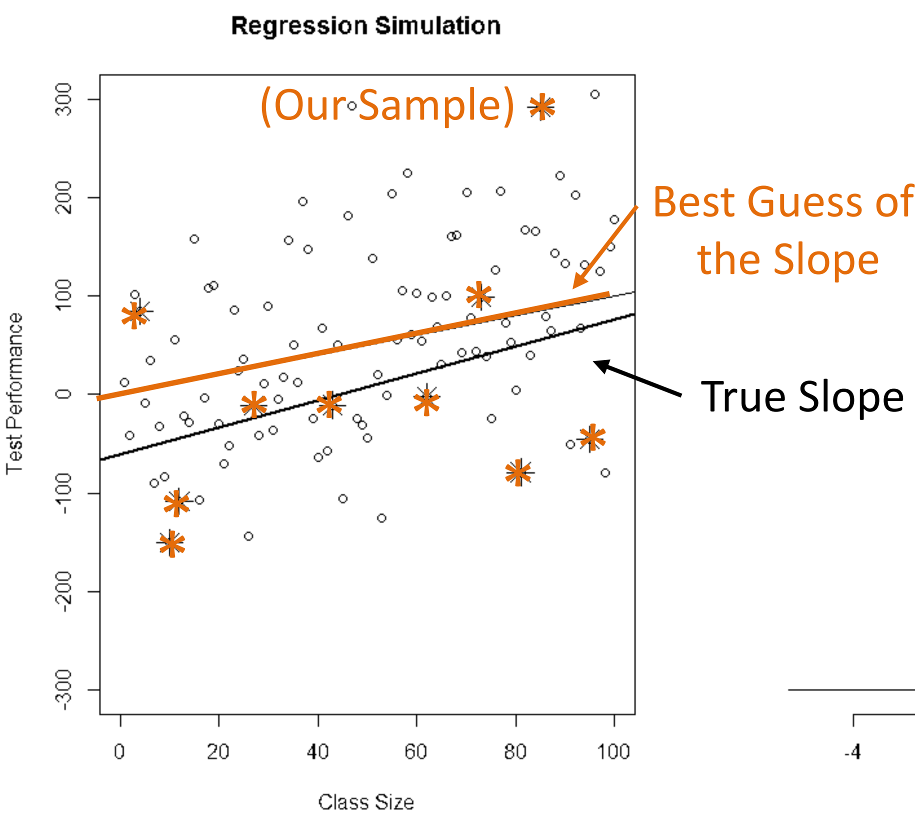
SAMPLE SIZE = 10

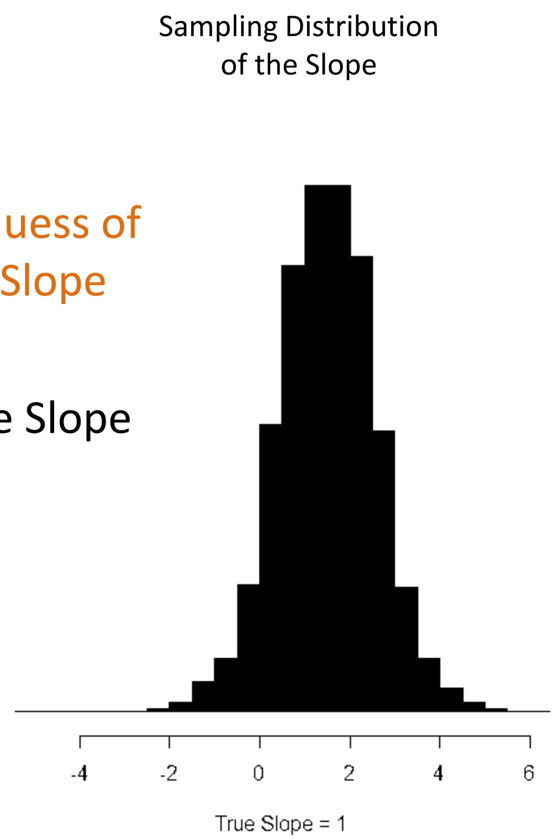
STANDARD ERROR OF THE SLOPE



SAMPLE SIZE = 50

STANDARD ERROR OF THE SLOPE





$$SE_{b_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$$

"AVERAGE ERROR" (OF THE SLOPE ESTIMATE)

THE INTUITIVE STANDARD ERROR

$$SE_{b_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$$

NOTE:

$$\operatorname{var}(x) = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \Rightarrow$$

$$(n-1) \cdot \operatorname{var}(x) = \sum (x_i - \bar{x})^2$$

THUS:

$$SE_{b_1} = \sqrt{\frac{\sigma_\varepsilon^2}{(n-1)\operatorname{var}(x)}}$$

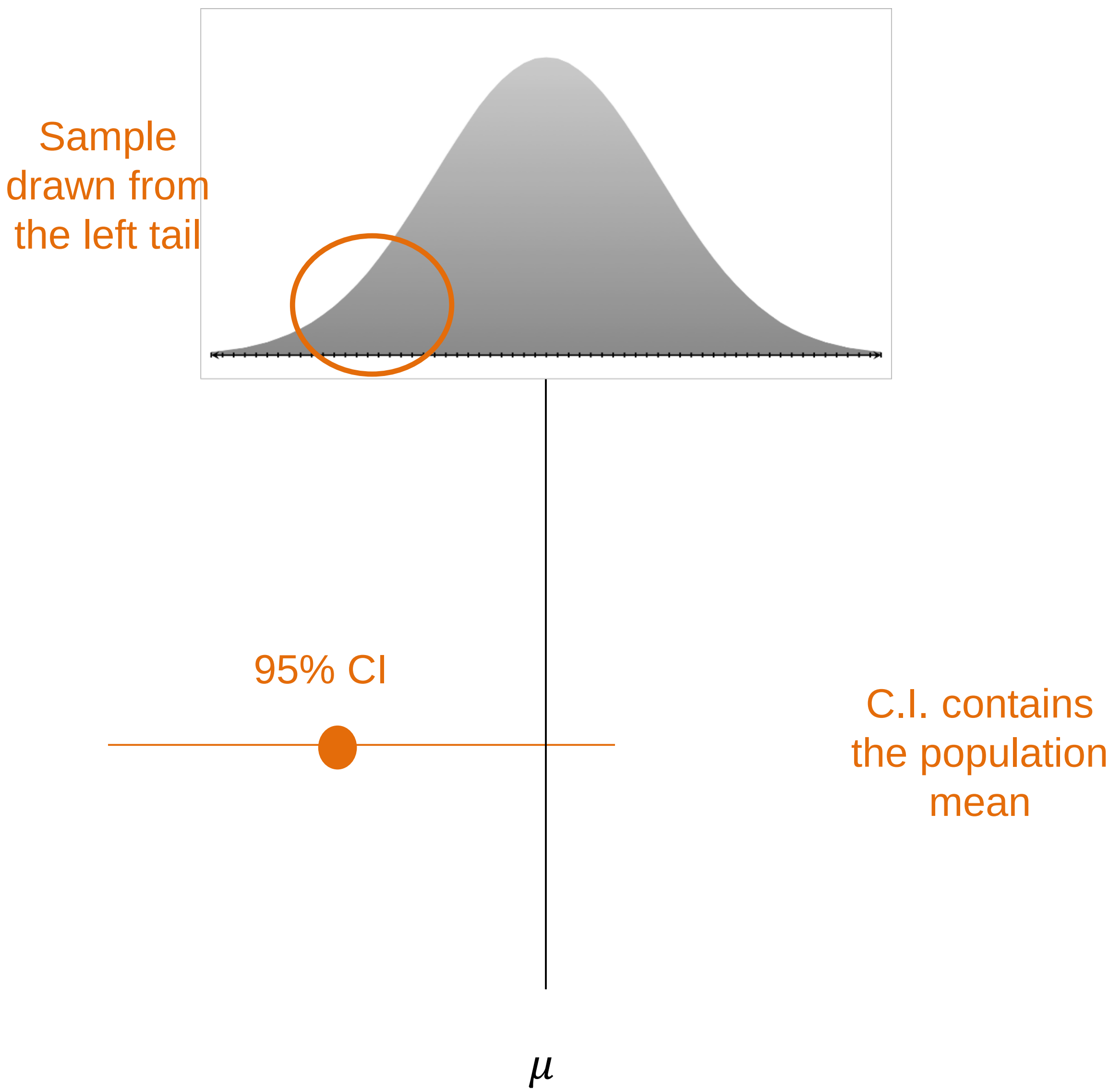
THEREFORE:

$$SE_{b_1} \approx \frac{residual_y}{sample\,size \cdot \operatorname{var}(x)}$$

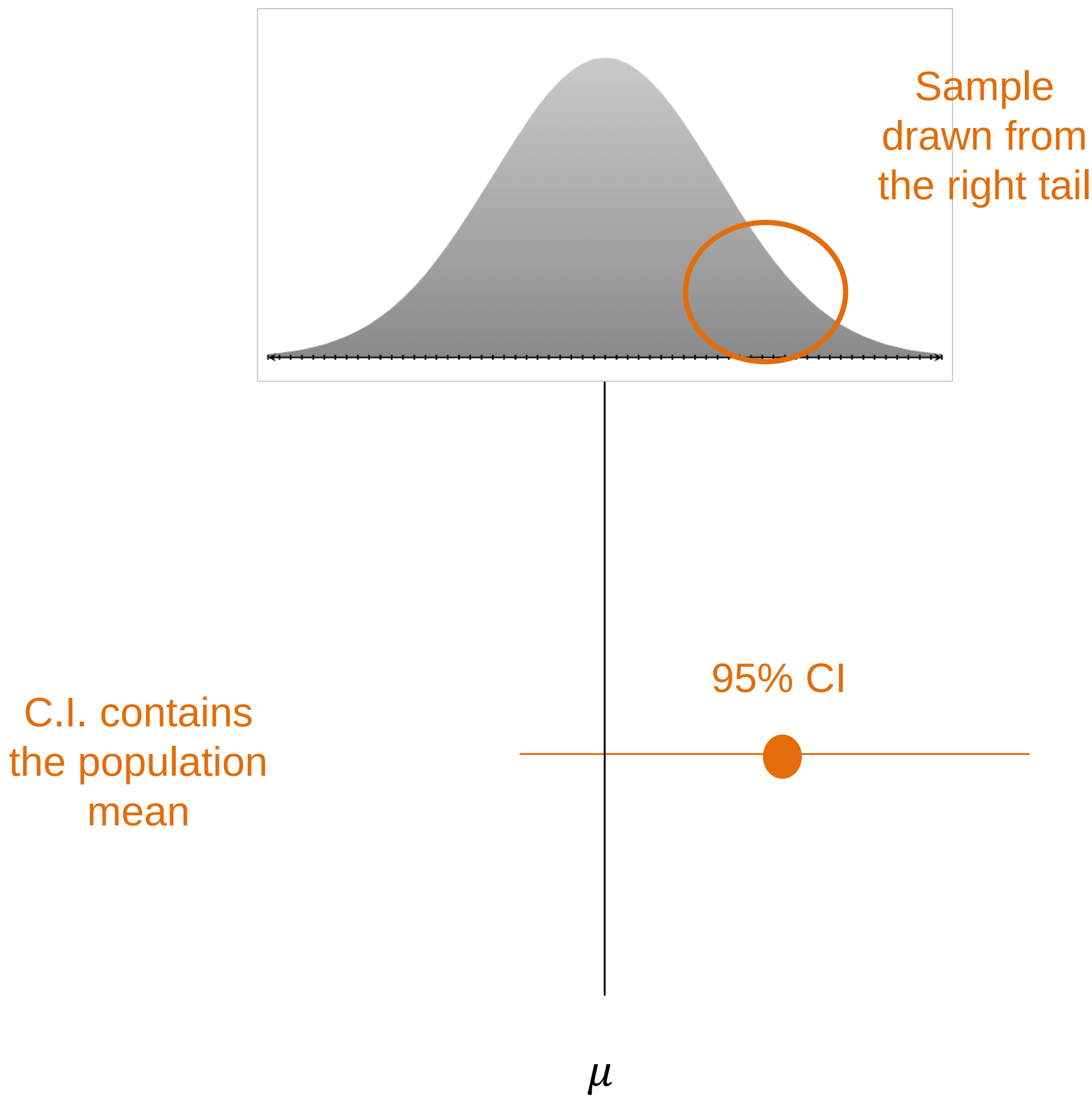
Confidence Intervals

An interval that will contain the true slope in 95% of the samples that we draw.

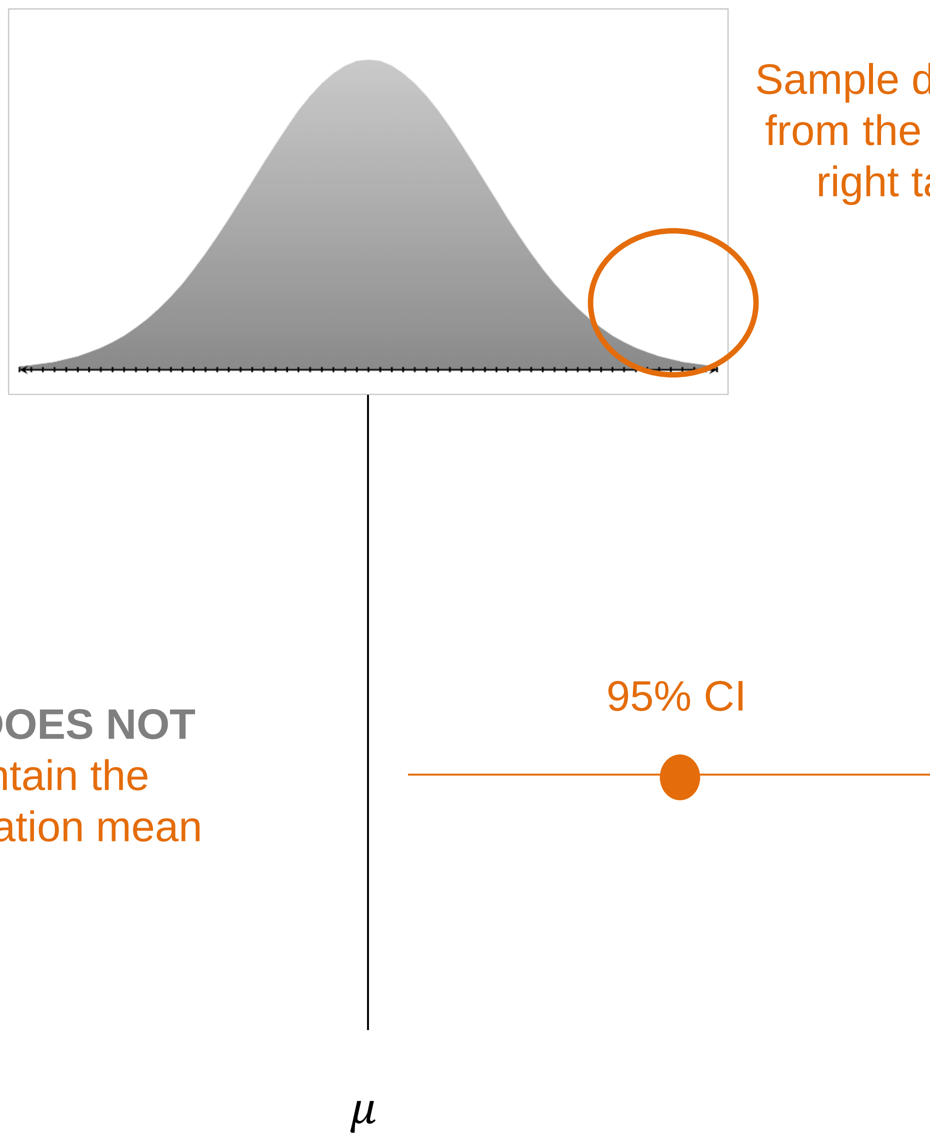
CONFIDENCE INTERVALS



CONFIDENCE INTERVALS



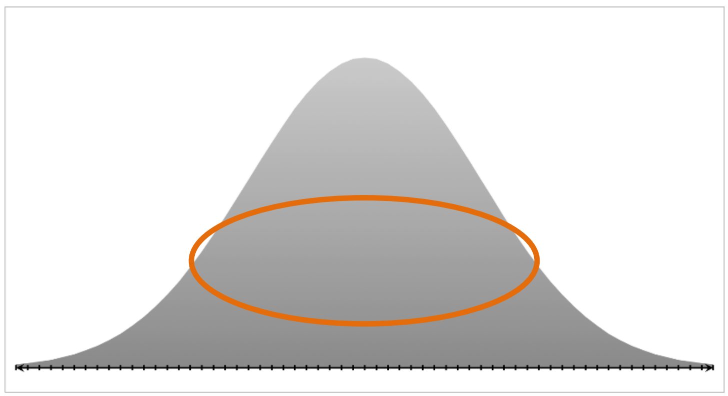
CONFIDENCE INTERVALS

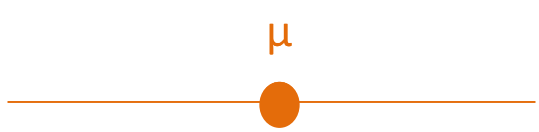


C.I. **DOES NOT** contain the population mean

How often will this happen?

CONFIDENCE INTERVALS

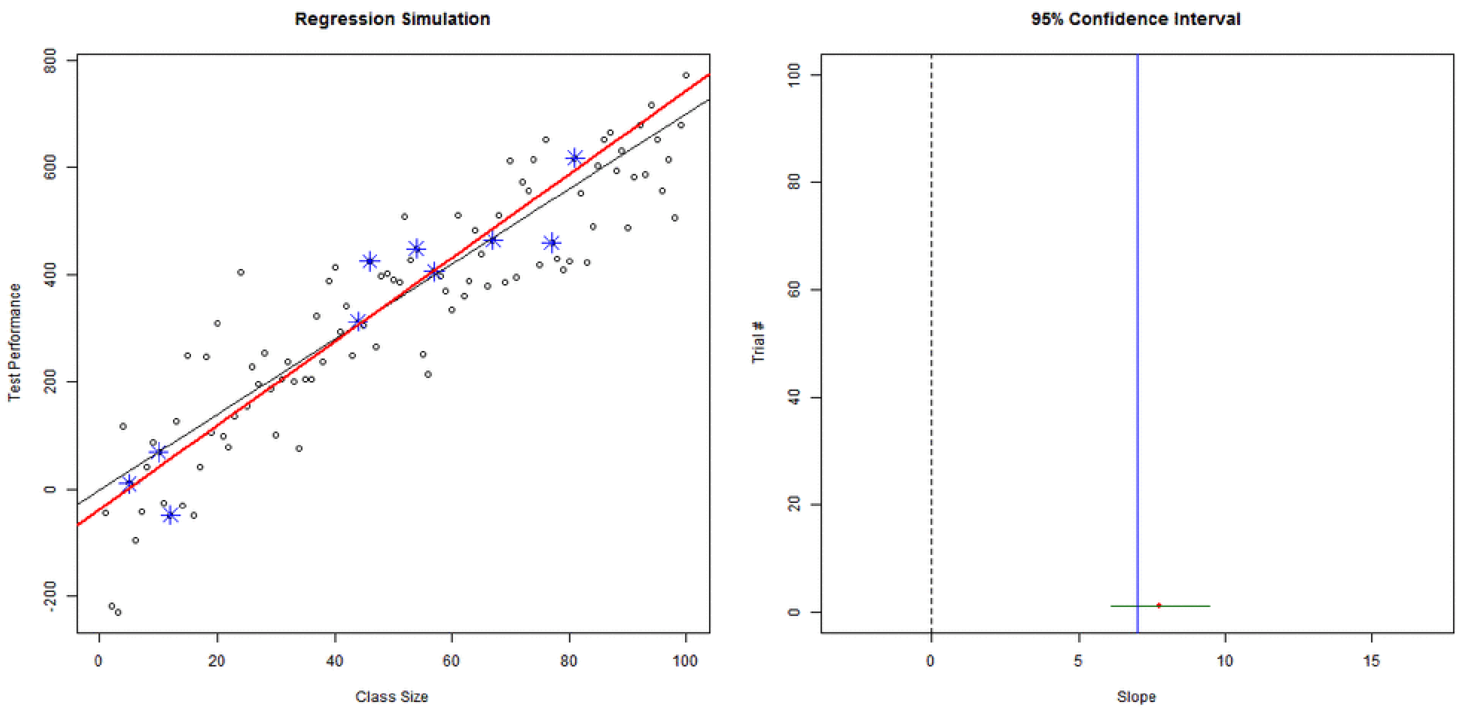




$$\bar{x} - t \cdot SE_{\bar{x}} < \mu > \bar{x} + t \cdot SE_{\bar{x}}$$

C.I. of the sample mean

CONFIDENCE INTERVAL OF THE SLOPE



$$b_1 - t \cdot SE_{b_1} < \beta_1 > b_1 + t \cdot SE_{b_1}$$

C.I. of the Slope

THE ROAD MAP

	Of the Mean:	Of the Slope:
Variance:	$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ (for x)	$\sigma_\varepsilon^2 = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}$ (using the residual)
↓		
Standard Deviation:	$\sigma_x = \sqrt{\sigma_x^2}$	$\sigma_\varepsilon = \sqrt{\sigma_\varepsilon^2}$
↓		
Standard Error:	$SE_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$	$SE_{b_1} = \sqrt{\frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$
↓		
Confidence Interval	$\mu = \bar{x} \pm t \cdot SE_{\bar{x}}$ (of the mean)	$\beta_1 = b_1 \pm t \cdot SE_{b_1}$ (of the slope)

All of the statistical concepts that you have learned in the previous course using variance, standard errors, and confidence intervals of a estimates of the mean from a single variable apply to regression, but they have to be adapted.

Make note that statistical concepts always need to be followed by the phrase "of the" because they are general concepts and the specific calculations are determined by the variables you are working with. The standard error around an estimated mean is different than the standard error around an estimated slope.

Understanding regression error and the standard error of the regressors.

What should be clear in my mind?

1. **What is a sampling distribution?**
2. **What is the relationship between the sampling distribution and the standard error?**
3. **We care about the sampling variance of which statistic in regression?**
4. **What role does the standard error play in the confidence interval?**