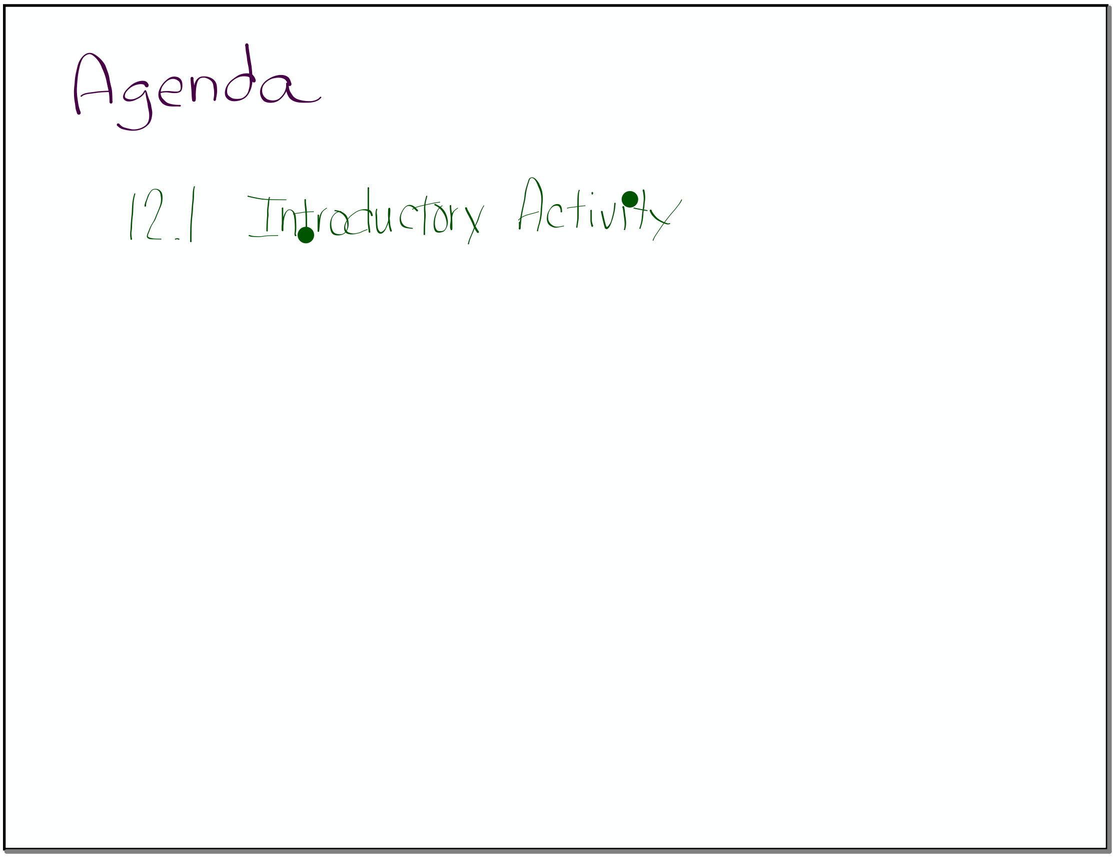
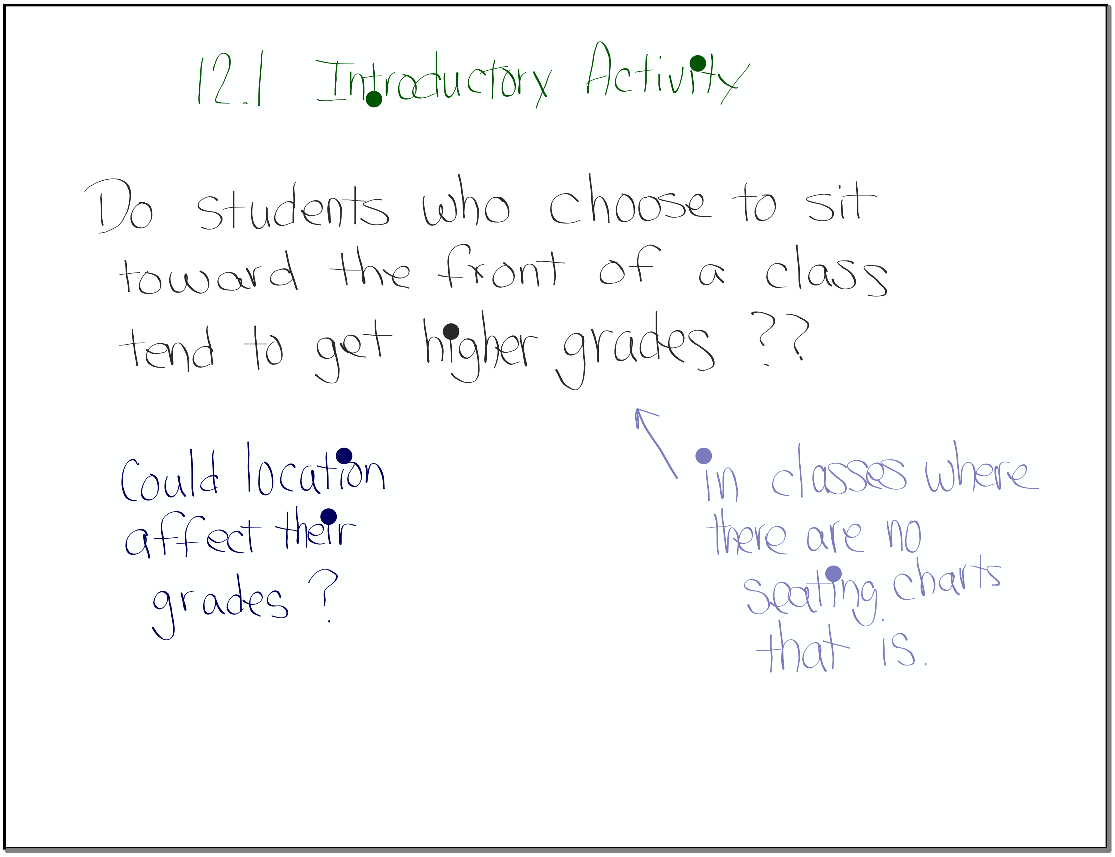


Agenda

12.1 Introductory Activity



12.1 Introductory Activity

Do students who choose to sit toward the front of a class tend to get higher grades ??

Could location affect their grades ?

in classes where there are no seating charts that is.

Data from a class
in Michigan is on
the handout.

Lesson 12.1-Day 1: **Does seat location matter? Part 1**

A teacher in Michigan randomly assigned students to rows last year. At the end of the time period that the seats were in place she analyzed grades. Did students who sit in the front rows do better than students who sit farther away?

Row	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3
Score	76	77	94	99	88	90	83	85	74	79	77	79	90	88	68	78	83	79

Row	4	4	4	4	4	4	5	5	5	5	5	5
Score	94	72	101	70	63	76	76	65	67	96	79	96

work through the
front side.

Row	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3
Score	76	77	94	99	88	90	83	85	74	79	77	79	90	88	68	78	83	79

Row	4	4	4	4	4	4	5	5	5	5	5	5
Score	94	72	101	70	63	76	76	65	67	96	79	96

1. Is this an observational study or an experiment? Why?

2. Why is it important to randomly assign the students to seats rather than letting each student choose his or her own seat?

3. How many variables are we measuring?______ Are they categorical or quantitative?

What is the explanatory variable (x)?___________________ Response variable(y)?_______________

Row	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3
Score	76	77	94	99	88	90	83	85	74	79	77	79	90	88	68	78	83	79

Row	4	4	4	4	4	4	5	5	5	5	5	5
Score	94	72	101	70	63	76	76	65	67	96	79	96

1. Is this an observational study or an experiment? Why?

Experiment - a treatment (row) is imposed.

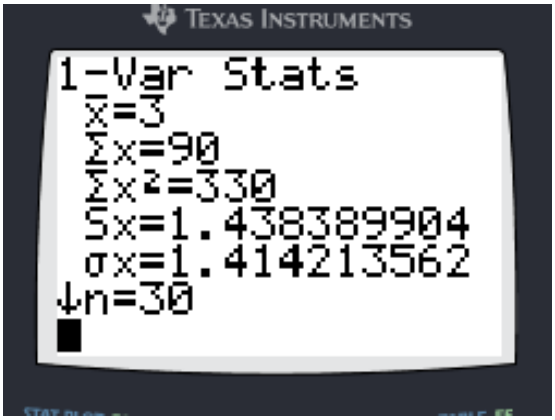
2. Why is it important to randomly assign the students to seats rather than letting each student choose his or her own seat?

It allow us to show causation.

3. How many variables are we measuring?_2_ Are they categorical or [✓ quantitative?]

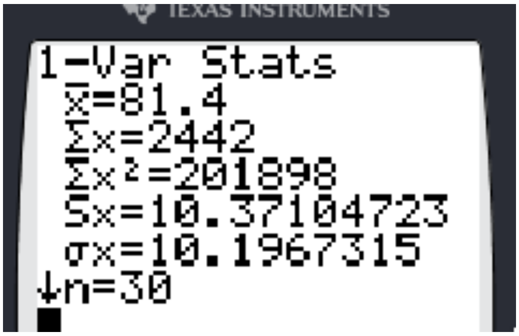
What is the explanatory variable (x)?_Row_ Response variable(y)?_Score_

L₁



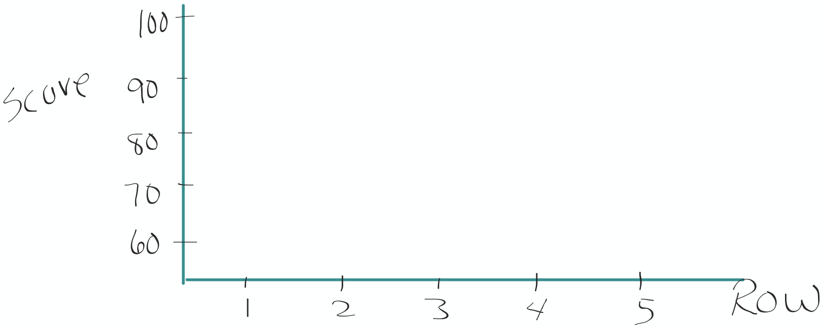
```
1-Var Stats
 x̄=3
 Σx=90
 Σx²=330
 Sx=1.438389904
 σx=1.414213562
↓n=30
```

L₂



```
1-Var Stats
 x̄=81.4
 Σx=2442
 Σx²=201898
 Sx=10.37104723
 σx=10.1967315
↓n=30
```

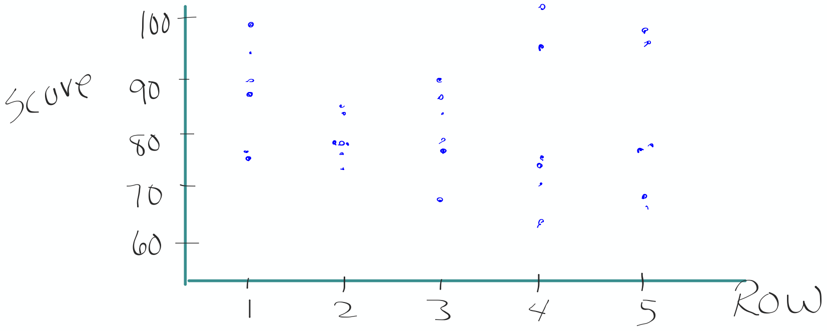
4. Use stapplet.com or your GDC to make a scatterplot. Sketch it below.



5. Find the least squares regression line (LSRL):____

6. What is the slope of the LSRL:__________ Interpret the slope in the context of the problem.

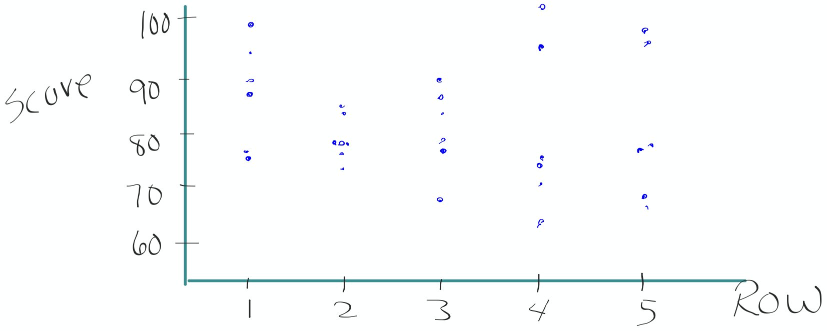
4. Use stapplet.com or your GDC to make a scatterplot. Sketch it below.



5. Find the least squares regression line (LSRL): $\widehat{Score} = 85.95 - 1.517(Raw)$

6. What is the slope of the LSRL: -1.517 Interpret the slope in the context of the problem.

4. Use stapplet.com or your GDC to make a scatterplot. Sketch it below.



5. Find the least squares regression line (LSRL): $\widehat{Score} = 85.95 - 1.517(Raw)$

6. What is the slope of the LSRL: -1.517 Interpret the slope in the context of the problem.

increase of 1 row

For each additional row, the predicted score goes down by 1.517 points.

We need to find out how likely it is that our slope occurs purely by chance if row has no effect on exam score time for simulation

↓

We'll create a sampling distribution to find out.

7. Does the negative slope provide convincing evidence that sitting closer causes higher achievement, or is it plausible that the association is purely by chance because of random assignment?

In order to answer this question, we need to know more about "*purely by chance because of random assignment*". If we assume that seat location has No effect on Exam Score, then we could just randomly assign all 30 Exam Scores to each of the seat locations. We will do this by writing down each of the 30 Exam Scores onto an index card, shuffle the index cards, and then randomly assign them to seat locations.

In pairs, shuffle up the note cards and randomly assign 6 students into each of the 5 rows. Record the results:

~~66~~ → 99

~~89~~ → 68

Row 1: _____, _____, _____, _____, _____, _____

Row 2: _____, _____, _____, _____, _____, _____

Row 3: _____, _____, _____, _____, _____, _____

Row 4: _____, _____, _____, _____, _____, _____

Row 5: _____, _____, _____, _____, _____, _____ Now find the slope of the LSRL:_ b= __________

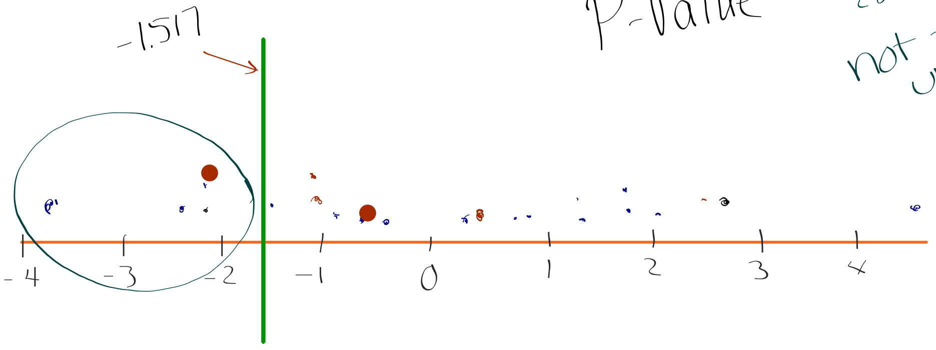
$\hat{y} = a + \underbrace{b}_{\text{slope}}x$

You have now calculated three different possible values for the slope based on random assignment. Take these 3 values to the dotplot on the whiteboard in the front of the room. When everyone in class has recorded their data, copy the dotplot below:

Sampling Distribution of Slopes

P-value $= \frac{5}{28} \approx 17.9\%$

not too unusual



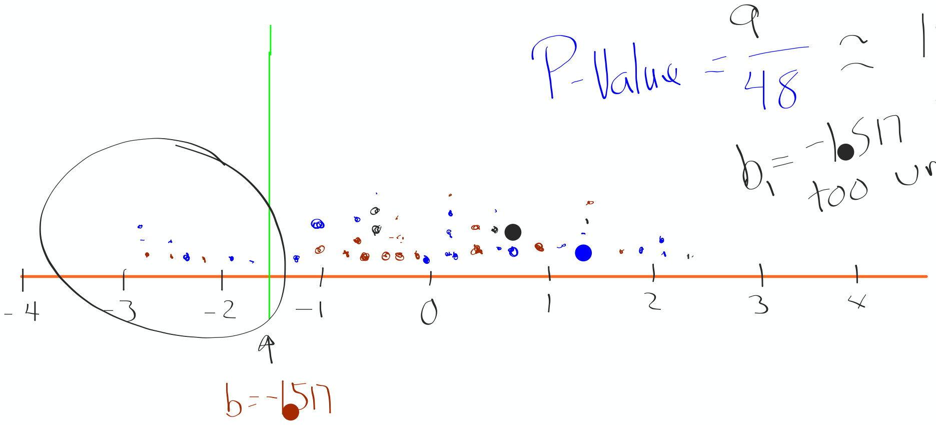
Different samples of the same size from the same population will give different estimates for the slope

You have now calculated three different possible values for the slope based on random assignment. Take these 3 values to the dotplot on the whiteboard in the front of the room. When everyone in class has recorded their data, copy the dotplot below:

Sampling Distribution of Slopes

P-Value $= \frac{9}{48} \approx 18.75\%$

$b_1 = -1.517$ is not too unusual

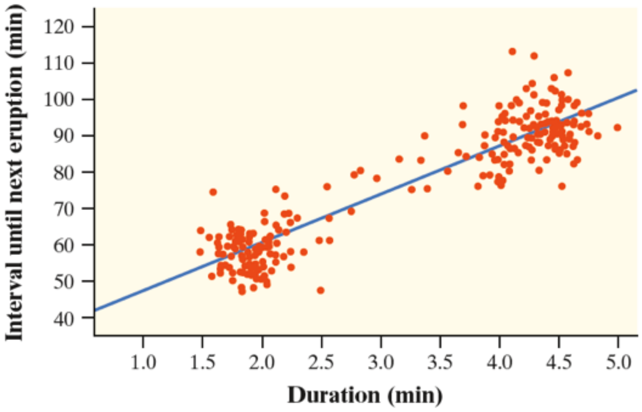


Learning Targets

✓CHECK the conditions for performing inference about the slope β_1 of the population (true) regression line.

✓INTERPRET the values of β_0, β_1, σ, and SE_{b_1} in context, and DETERMINE these values from computer output.

Inference for Linear Regression

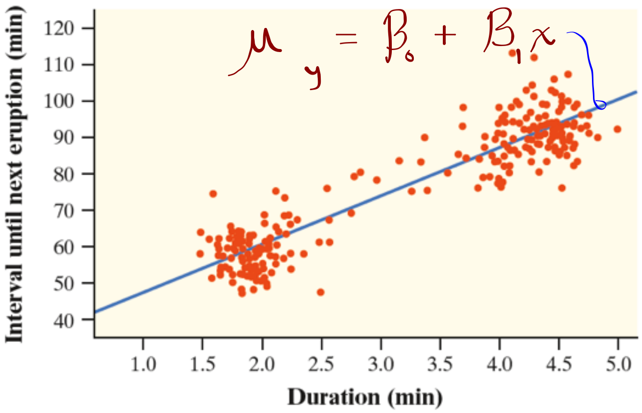


Scatterplot of the duration and interval between eruptions of Old Faithful for all 263 eruptions in a single month. The **population least-squares regression line** is shown in blue.

A regression line calculated from every value in the population is called a **population regression line** (true regression line). The equation of a population regression line is $\mu_y = \beta_0 + \beta_1 x$ where

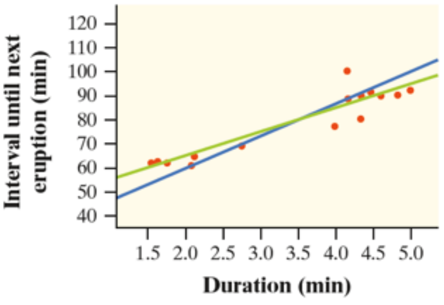
- μ_y is the mean *y*-value for a given value of *x*.
- β_0 is the population *y* intercept.
- β_1 is the population slope.

Inference for Linear Regression

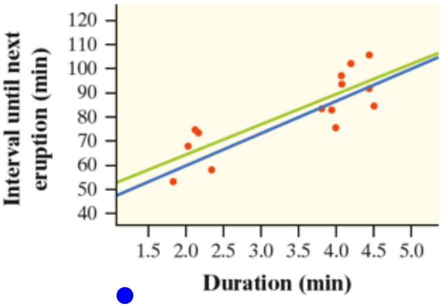


Scatterplot of the duration and interval between eruptions of Old Faithful for all 263 eruptions in a single month. The **population least-squares regression line** is shown in blue.

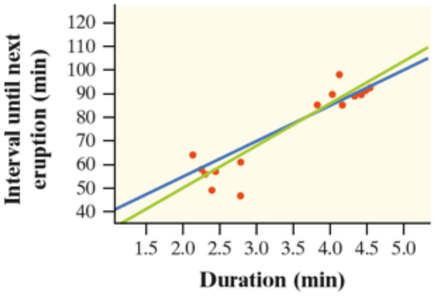
Here are scatterplots and **least-squares regression lines (in green)** for three different SRSs of 15 Old Faithful eruptions, along with the **population regression line (in blue)**.



Sample 1: $\hat{y} = 44 + 10.0x$



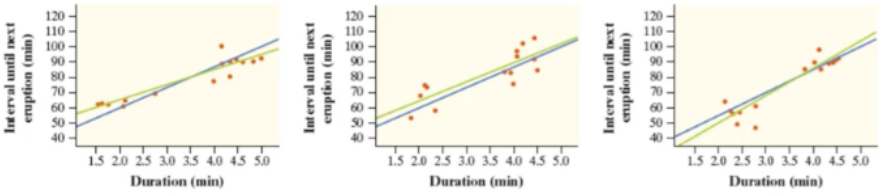
Sample 2: $\hat{y} = 39 + 12.5x$



Sample 3: $\hat{y} = 24 + 15.7x$

Here are scatterplots and **least-squares regression lines (in green)** for three different SRSs of 15 Old Faithful eruptions, along with the **population regression line (in blue)**.

Different samples of the same size from the same population will give different estimates for the slope.



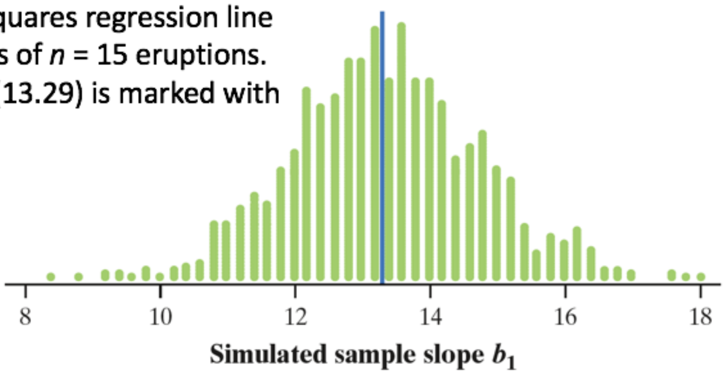
Sample 1: $\hat{y} = 44 + 10.0x$

Sample 2: $\hat{y} = 39 + 12.5x$

Sample 3: $\hat{y} = 24 + 15.7x$

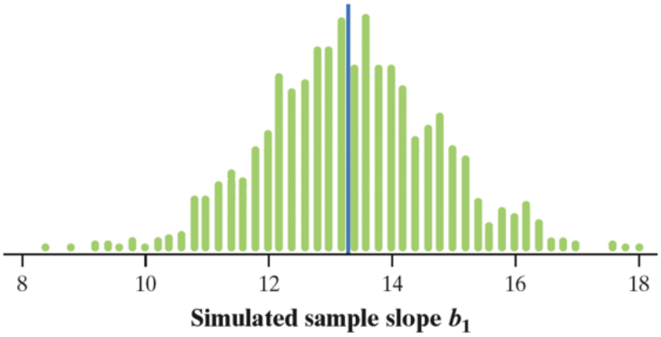
The pattern is described by its sampling distribution

The figure shows a dotplot of the sample slope b_1 of the least-squares regression line in 1000 simulated SRSs of $n = 15$ eruptions. The population slope (13.29) is marked with a **blue vertical line**.

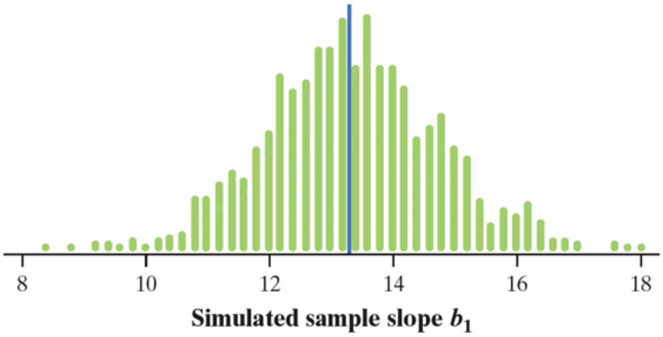


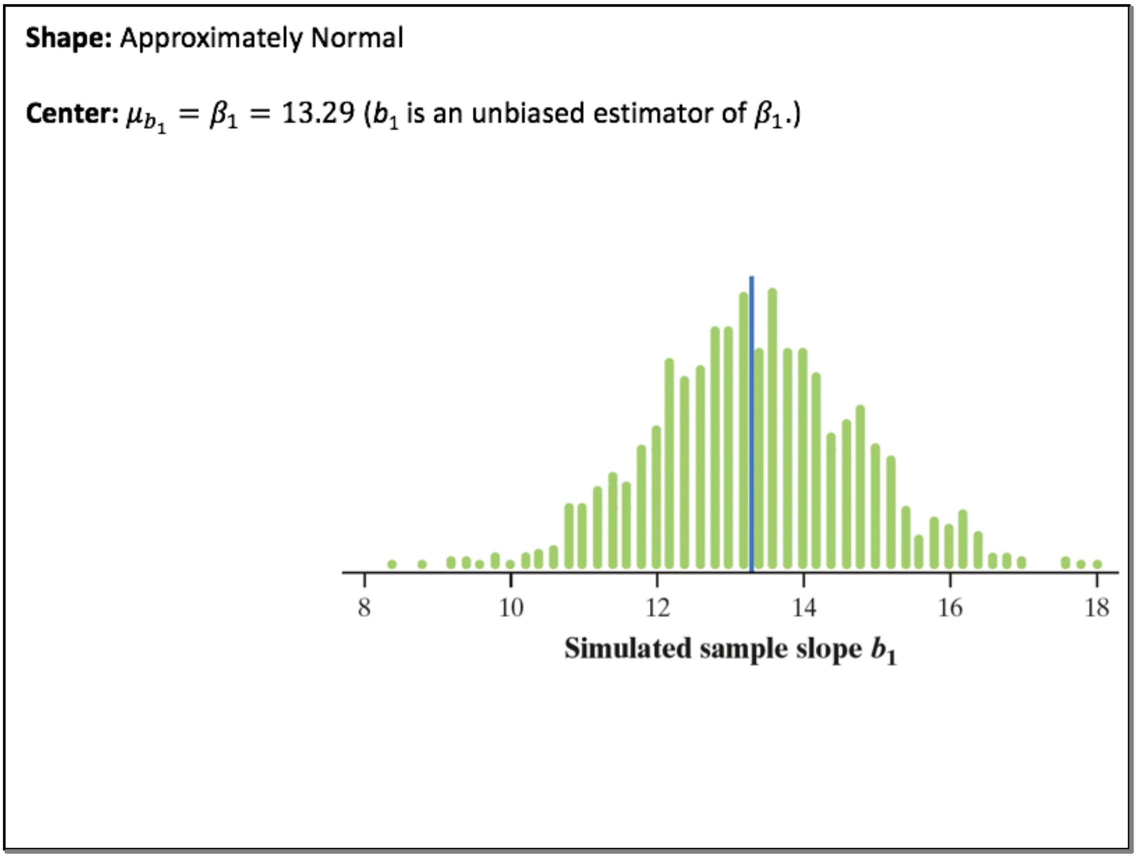
Sampling Distribution of b_1

which estimates the population slope β_1



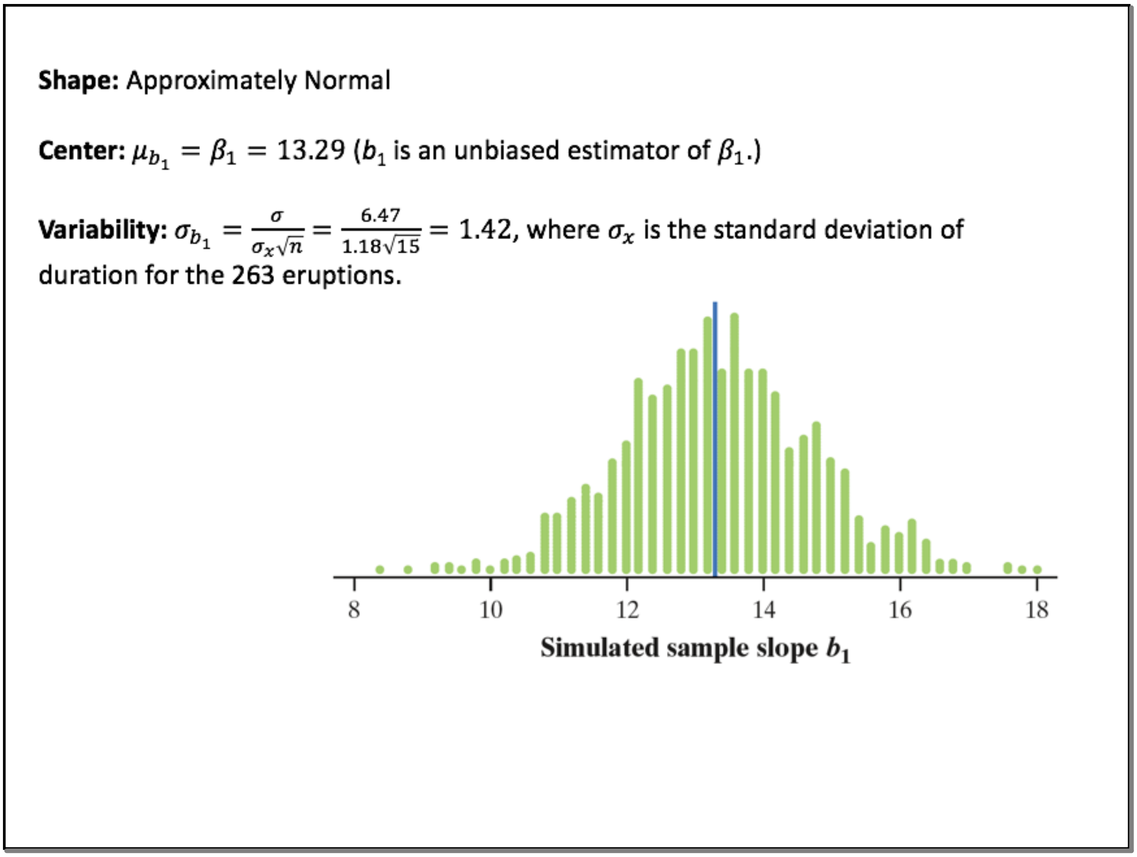
Shape: Approximately Normal





Shape: Approximately Normal

Center: $\mu_{b_1} = \beta_1 = 13.29$ (b_1 is an unbiased estimator of β_1.)

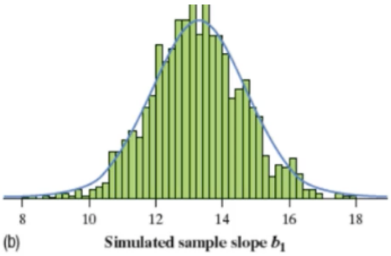


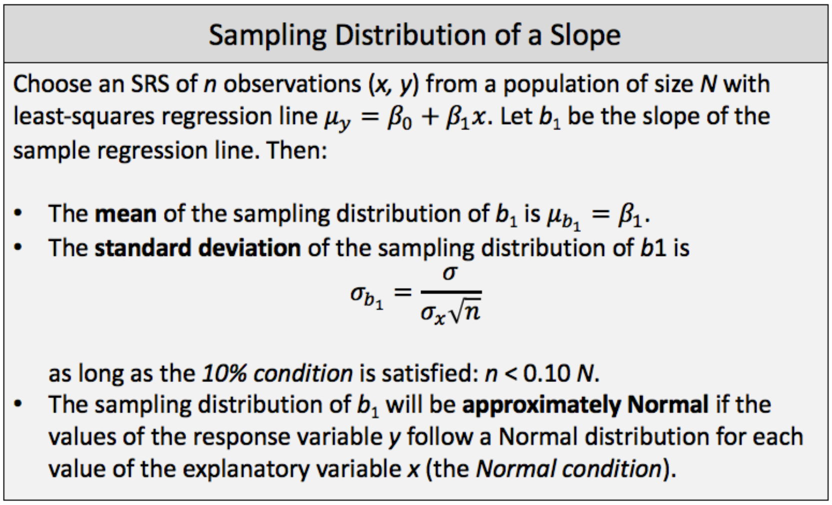
Shape: Approximately Normal

Center: $\mu_{b_1} = \beta_1 = 13.29$ (b_1 is an unbiased estimator of β_1.)

Variability: $\sigma_{b_1} = \frac{\sigma}{\sigma_x \sqrt{n}} = \frac{6.47}{1.18\sqrt{15}} = 1.42$, where σ_x is the standard deviation of duration for the 263 eruptions.

When certain conditions are met, we can anticipate the shape, center, and variability of the sampling distribution of the sample slope.





Sampling Distribution of a Slope

Choose an SRS of *n* observations (*x, y*) from a population of size *N* with least-squares regression line $\mu_y = \beta_0 + \beta_1 x$. Let b_1 be the slope of the sample regression line. Then:

- The **mean** of the sampling distribution of b_1 is $\mu_{b_1} = \beta_1$.
- The **standard deviation** of the sampling distribution of *b*1 is

$$\sigma_{b_1} = \frac{\sigma}{\sigma_x\sqrt{n}}$$

  as long as the *10% condition* is satisfied: $n < 0.10\,N$.
- The sampling distribution of b_1 will be **approximately Normal** if the values of the response variable *y* follow a Normal distribution for each value of the explanatory variable *x* (the *Normal condition*).

	We Used	to estimate
Ch. 8	$\hat{p}$	p
	$\bar{x}$	μ
Ch. 10	$(\hat{p}_1 - \hat{p}_1)$	$(p_1 - p_2)$
	$(\bar{x}_1 - \bar{x}_2)$	$(\mu_1 - \mu_2)$
Ch. 12	$\hat{y} = a + bx$ sample regression line	$\mu_y = \beta_0 + \beta_1 x$ population regression line

we'll focus on estimating slope β_1

Big Idea: If the data come from a random sample or randomized experiment, the least-squares regression line is just an *estimate* of the population (true) least-squares regression line.

- Population line: $\mu_y = \beta_0 + \beta_1 x$
- Sample line: $\hat{y} = b_0 + b_1 x$

Lesson 12.1: Day 1: Sampling Distribution of *b_1*

Write small ☺

Lesson 12.1: Day 1: Sampling Distribution of *b_1*

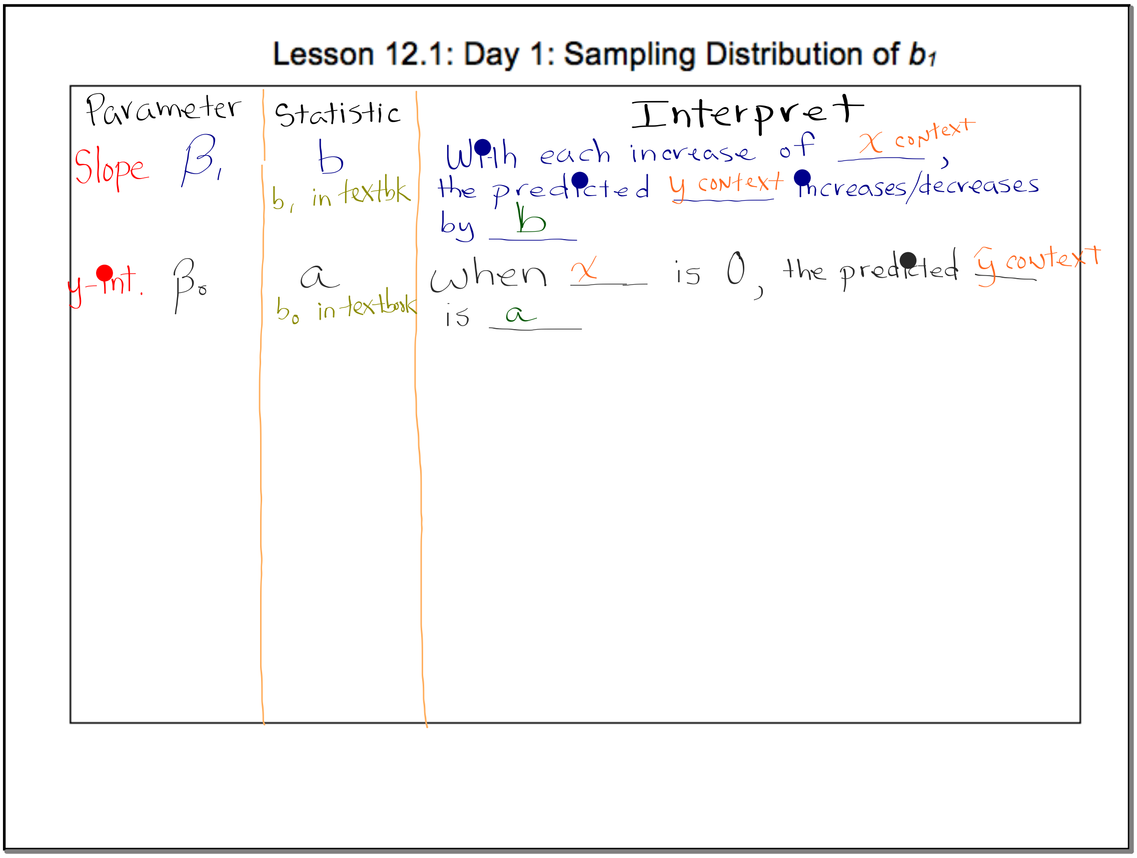
Parameter	Statistic	Interpret

Lesson 12.1: Day 1: Sampling Distribution of b_1

Parameter	Statistic	Interpret
Slope β_1	b	With each increase of x context, the predicted y context increases/decreases by b

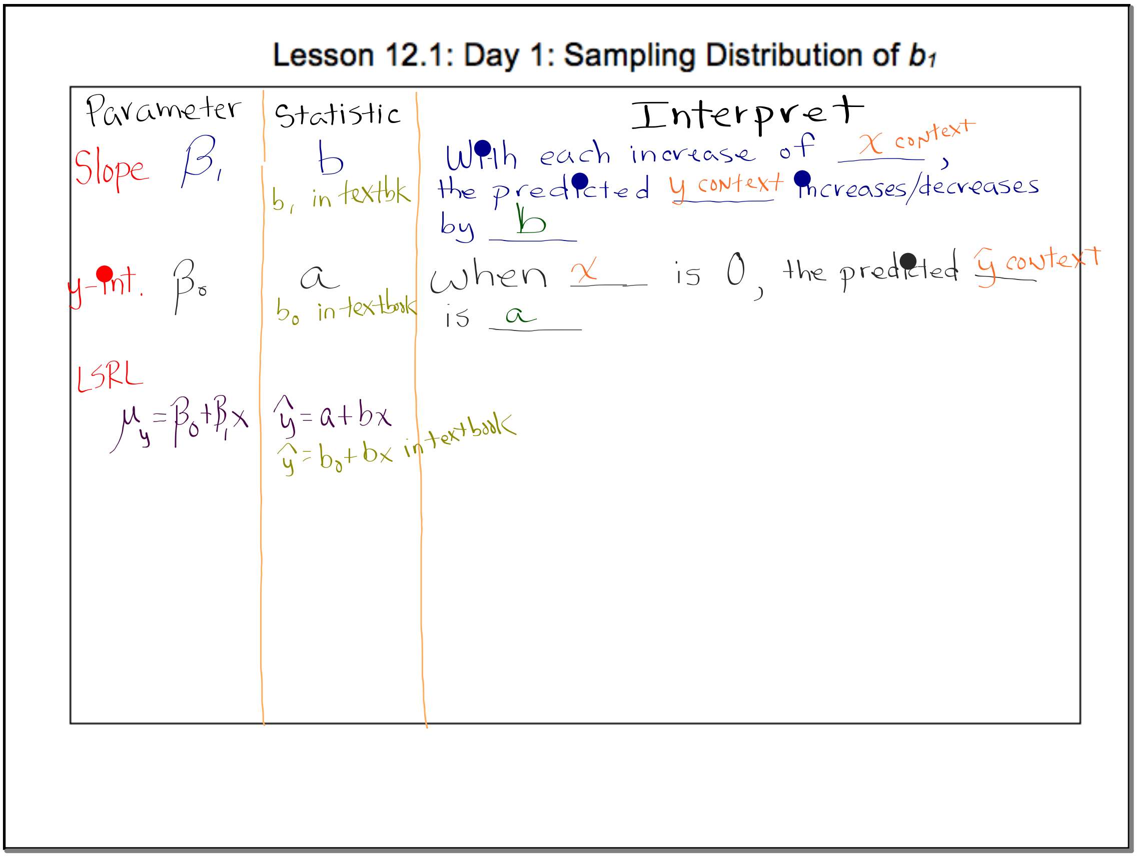
Lesson 12.1: Day 1: Sampling Distribution of b_1

Parameter	Statistic	Interpret
Slope β_1	b	With each increase of x context, the predicted y context increases/decreases by b
y-int. β_0	a	when x is 0, the predicted y context is a



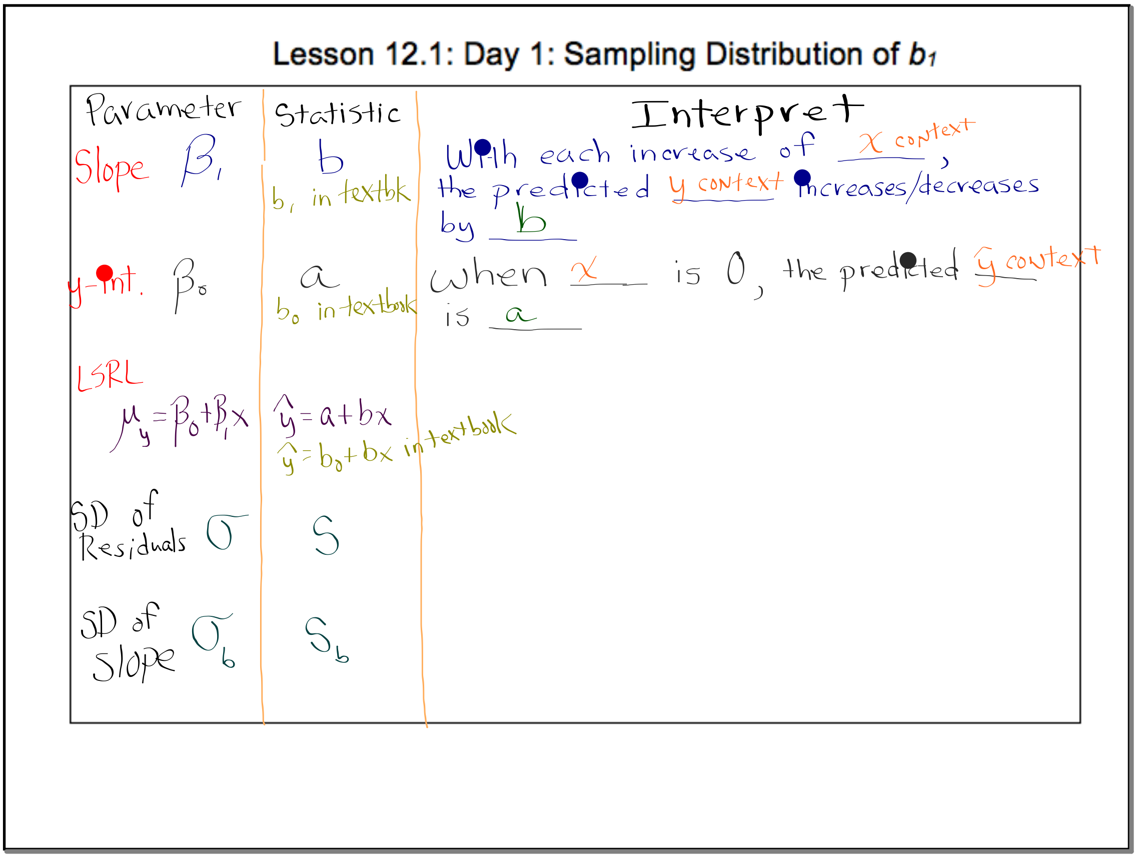
Lesson 12.1: Day 1: Sampling Distribution of b_1

Parameter	Statistic	Interpret
Slope β_1	b b_1 in textbk	With each increase of x context, the predicted y context increases/decreases by b
y-int. β_0	a b_0 in textbook	when x is 0, the predicted $\hat{y}$ context is a



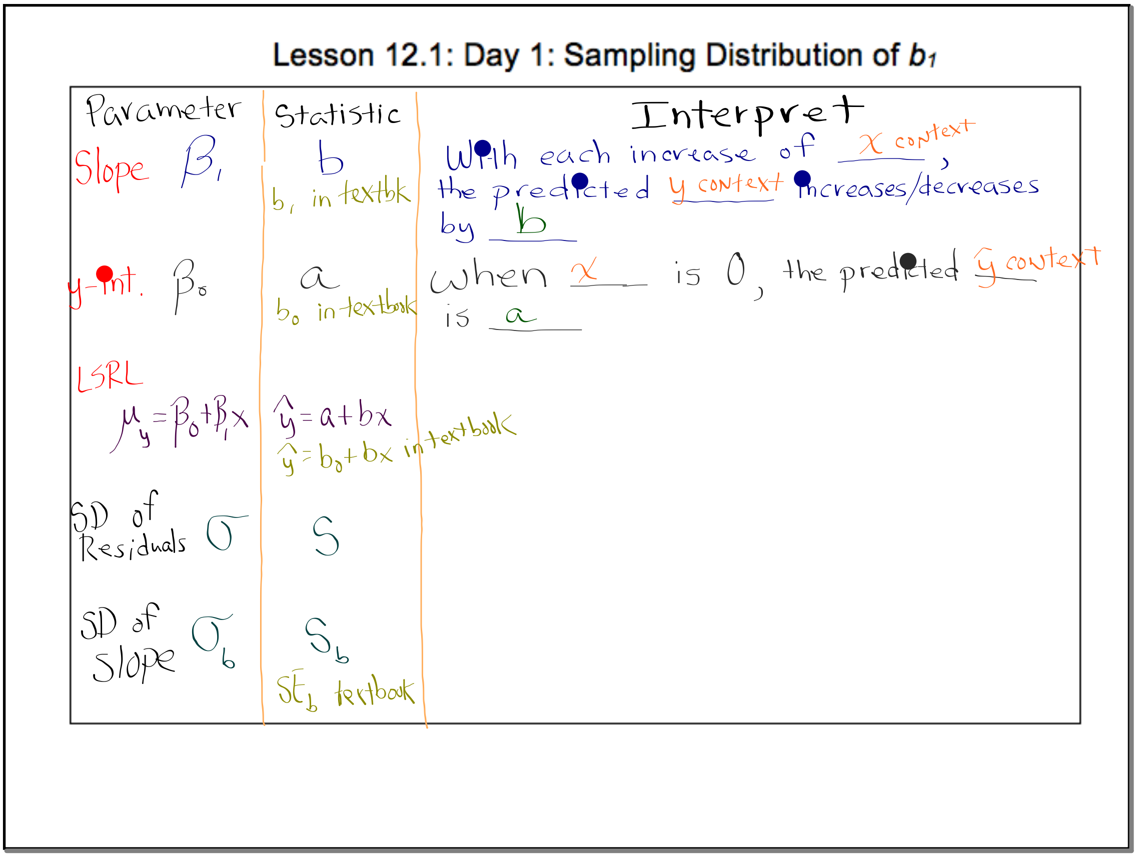
Lesson 12.1: Day 1: Sampling Distribution of b_1

Parameter	Statistic	Interpret
Slope β_1	b b_1 in textbk	With each increase of x context, the predicted y context increases/decreases by b
y-int. β_0	a b_0 in textbook	when x is 0, the predicted $\hat{y}$ context is a
LSRL $\mu_y=\beta_0+\beta_1x$	$\hat{y}=a+bx$ $\hat{y}=b_0+bx$ in textbook	



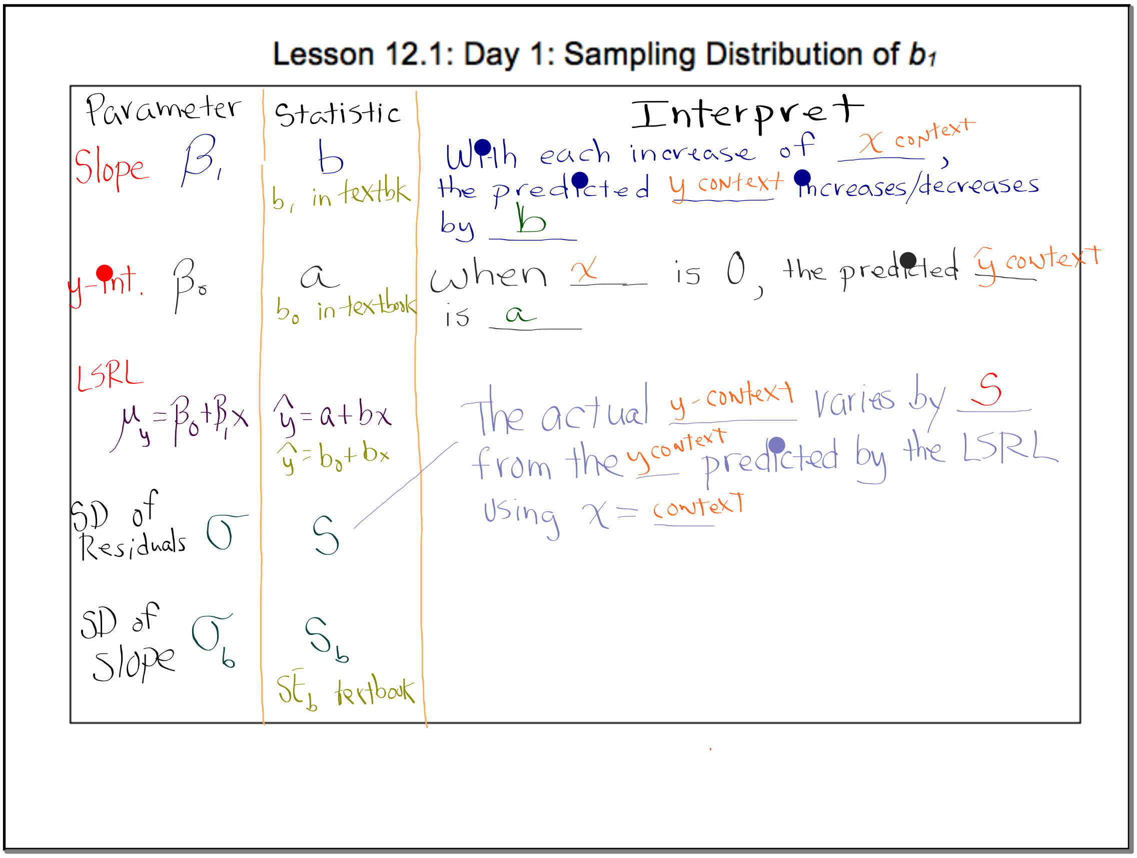
Lesson 12.1: Day 1: Sampling Distribution of *b₁*

Parameter	Statistic	Interpret
Slope β_1	b b_1 in textbk	With each increase of x context, the predicted y context increases/decreases by b
y-int. β_0	a b_0 in textbook	when x is 0, the predicted $\hat{y}$ context is a
LSRL $\mu_y=\beta_0+\beta_1 x$	$\hat{y}=a+bx$ $\hat{y}=b_0+bx$ in textbook	
SD of Residuals σ	s	
SD of Slope σ_b	s_b	



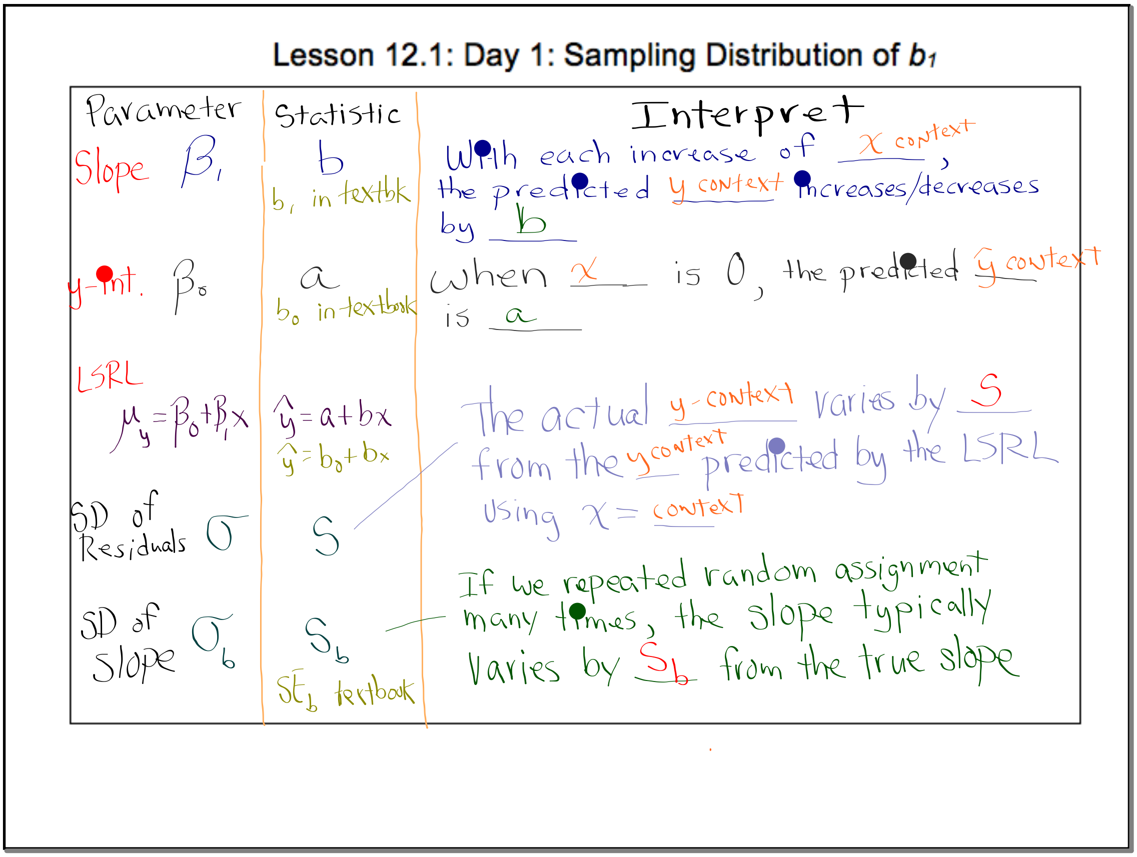
Lesson 12.1: Day 1: Sampling Distribution of *b₁*

Parameter	Statistic	Interpret
Slope β_1	b b_1 in textbk	With each increase of x context, the predicted y context increases/decreases by b
y-int. β_0	a b_0 in textbook	when x is 0, the predicted $\hat{y}$ context is a
LSRL $\mu_y=\beta_0+\beta_1 x$	$\hat{y}=a+bx$ $\hat{y}=b_0+bx$ in textbook	
SD of Residuals σ	s	
SD of Slope σ_b	s_b SE_b textbook	



Lesson 12.1: Day 1: Sampling Distribution of b_1

Parameter	Statistic	Interpret
Slope β_1	b b_1 in textbk	With each increase of x context, the predicted y context increases/decreases by b
y-int. β_0	a b_0 in textbk	when x is 0, the predicted $\hat{y}$ context is a
LSRL $\mu_y = \beta_0 + \beta_1 x$	$\hat{y} = a + bx$ $\hat{y} = b_0 + bx$	
SD of Residuals σ	s	The actual y-context varies by s from the y context predicted by the LSRL using x = context
SD of Slope σ_b	s_b SE_b textbook	

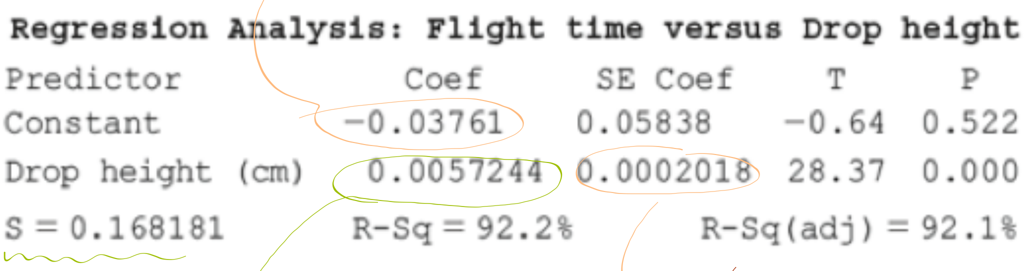


Lesson 12.1: Day 1: Sampling Distribution of b_1

Parameter	Statistic	Interpret
Slope β_1	b b_1 in textbk	With each increase of x context, the predicted y context increases/decreases by b
y-int. β_0	a b_0 in textbk	when x is 0, the predicted $\hat{y}$ context is a
LSRL $\mu_y = \beta_0 + \beta_1 x$	$\hat{y} = a + bx$ $\hat{y} = b_0 + bx$	
SD of Residuals σ	s	The actual y-context varies by s from the y context predicted by the LSRL using x = context
SD of Slope σ_b	s_b SE_b textbook	If we repeated random assignment many times, the slope typically varies by s_b from the true slope

(a) What is a, the estimate for β_0 (the true y-intercept)? Interpret this value.

y-intercept (a)



```
Regression Analysis: Flight time versus Drop height
Predictor            Coef        SE Coef        T       P
Constant          -0.03761      0.05838     -0.64   0.522
Drop height (cm)   0.0057244    0.0002018   28.37   0.000
S = 0.168181         R-Sq = 92.2%          R-Sq(adj) = 92.1%
```

Slope (b)

~~SE~~$_{b_1}$ S_b

(a) What is a, the estimate for β_0 (the true y-intercept)? Interpret this value.

$\cancel{b_0} = -0.03761$

$a =$

If the helicoptor is dropped from 0 cm, it is predicted to take −0.03761 seconds to land. This has no meaning.

(b) What is the estimate for β_1? Interpret this value.

$\cancel{b_1} = 0.0057244$

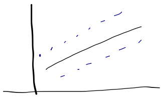
b

For every additional cm of drop height, the landing time rises by 0.0057 seconds

(c) What is the estimate for σ? Interpret this value.

$S = 0.16818$

the actual flight times typically vary by 0.168 seconds from predicted by LSRL.



(d) Give the standard error of the slope S_b. Interpret this value.

(c) What is the estimate for σ? Interpret this value.

$S = 0.16818$

the actual flight times typically vary by 0.168 seconds from predicted by LSRL.

(d) Give the standard error of the slope S_b. Interpret this value.

$S_b =$

$\cancel{SE_b} = 0.0002018$

If we repeated the random assignment many times, the slope of the sample LSRL typically varies by 0.0002 from the true slope.

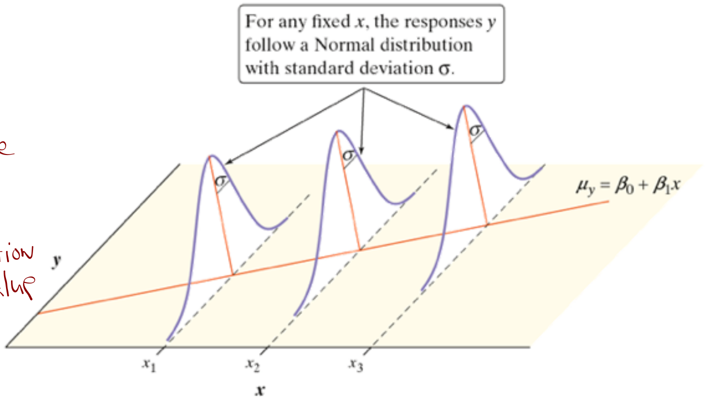
12.1...... 1, 7 and study the

conditions for inference on pp. 769-776

- if you have AP Stat T-shirt ideas please give to me (or let me know if you are planning to do so)

Conditions for Regression Inference

When the conditions for inference are met, the regression model looks like the one shown here. The line is the population (true) regression line, which shows how the mean response μ_y changes as the explanatory variable x changes. For any fixed value of x, the observed response y varies according to a Normal distribution having mean μ_y and standard deviation σ.

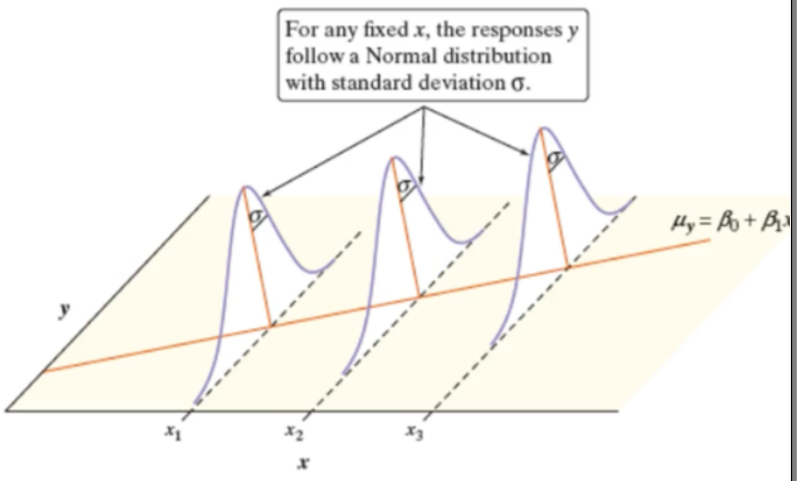

For any fixed x, the responses y follow a Normal distribution with standard deviation σ.

$\hat{y}$ sample

μ_y population y-value

- The conditions are based on the model for linear regression:

- **L**inear
- **I**ndependent
- **N**ormal
- **E**qual SD
- **R**andom



For any fixed x, the responses y follow a Normal distribution with standard deviation σ.

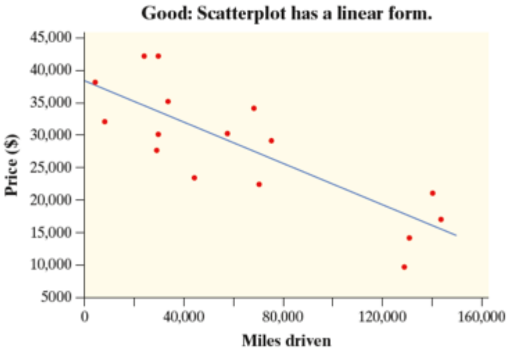
Conditions for Regression Inference

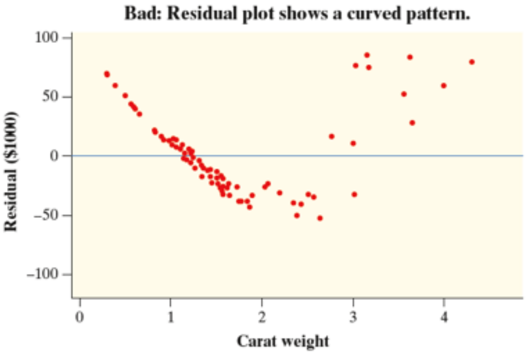
Suppose we have *n* observations on a quantitative explanatory variable *x* and a quantitative response variable *y.* Our goal is to study or predict the behavior of *y* for given values of *x.*

- **Linear:** The actual relationship between *x* and *y* is linear. For any fixed value of *x,* the mean response my falls on the population (true) regression line $\mu_y = \beta_0 + \beta_1 x$.
- **Independent:** Individual observations are independent of each other. When sampling without replacement, check the *10% condition.*
- **Normal:** For any fixed value of *x,* the response *y* varies according to a Normal distribution.
- **Equal SD:** The standard deviation of *y* (call it σ) is the same for all values of *x.*
- **Random:** The data come from a random sample from the population of interest or a randomized experiment.

Here’s a summary of how to check the conditions one by one.

Linear: Examine the scatterplot to see if the overall pattern is roughly linear. Make sure there are no leftover curved patterns in the residual plot.



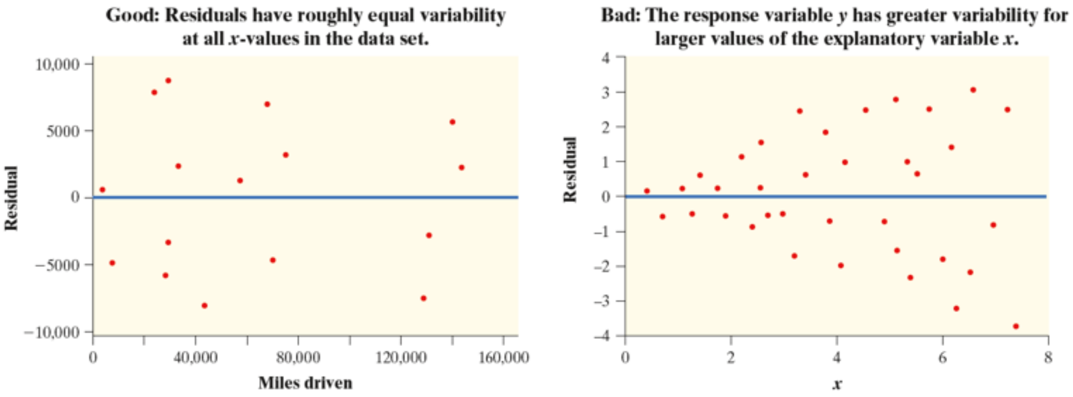


Here’s a summary of how to check the conditions one by one.

Independent: Knowing the value of the response variable for one individual shouldn’t help predict the value of the response variable for other individuals. If sampling is done without replacement, remember to check that the sample size is less than 10% of the population size (*10% condition*).

Normal: Make a histogram, dotplot, stemplot, boxplot, or Normal probability plot of the residuals and check for strong skewness or outliers. Ideally, we would check the Normality of the residuals at each possible value of *x*. Because we rarely have enough observations at each *x*-value, however, we make one graph of all the residuals to check for Normality.

Equal SD: Look at the scatter of the residuals above and below the "residual = 0" line in the residual plot. The variability of the residuals in the vertical direction should be roughly the same from the smallest to the largest *x*-value.


Good: Residuals have roughly equal variability at all *x*-values in the data set.

Bad: The response variable *y* has greater variability for larger values of the explanatory variable *x*.

The variability of the residuals in the vertical direction should be **ROUGHLY** the same as you scan across each of the x-values.

-Look for major violations only.

Random: See if the data came from a random sample from the population of interest or a randomized experiment. If not, we can't make inferences about a larger population or about cause and effect.

Check Your Understanding

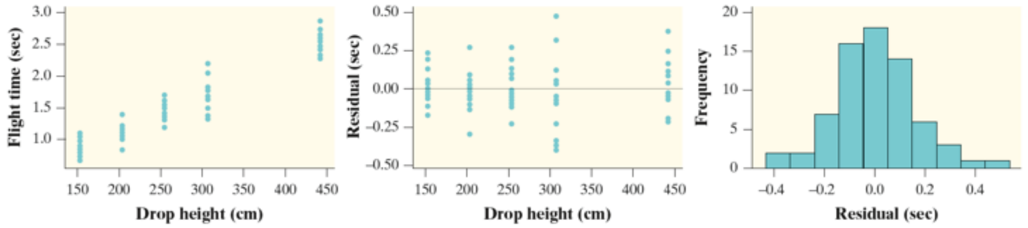
Mrs. Barrett's class did a fun experiment using paper helicopters. After making 70 helicopters using the same template, students randomly assigned 14 helicopters to each of five drop heights: 152 cm, 203 cm, 254 cm, 307 cm, and 442 cm.

Teams of students released the 70 helicopters in a random order and measured the flight times in seconds. The class used computer software to carry out a least-squares regression analysis for these data. Some output from this regression analysis is shown here. We checked conditions for performing inference earlier.

```
Regression Analysis: Flight time versus Drop height
Predictor           Coef        SE Coef      T       P
Constant          -0.03761     0.05838    -0.64   0.522
Drop height (cm)   0.0057244  0.0002018   28.37   0.000
S = 0.168181       R-Sq = 92.2%       R-Sq(adj) = 92.1%
```

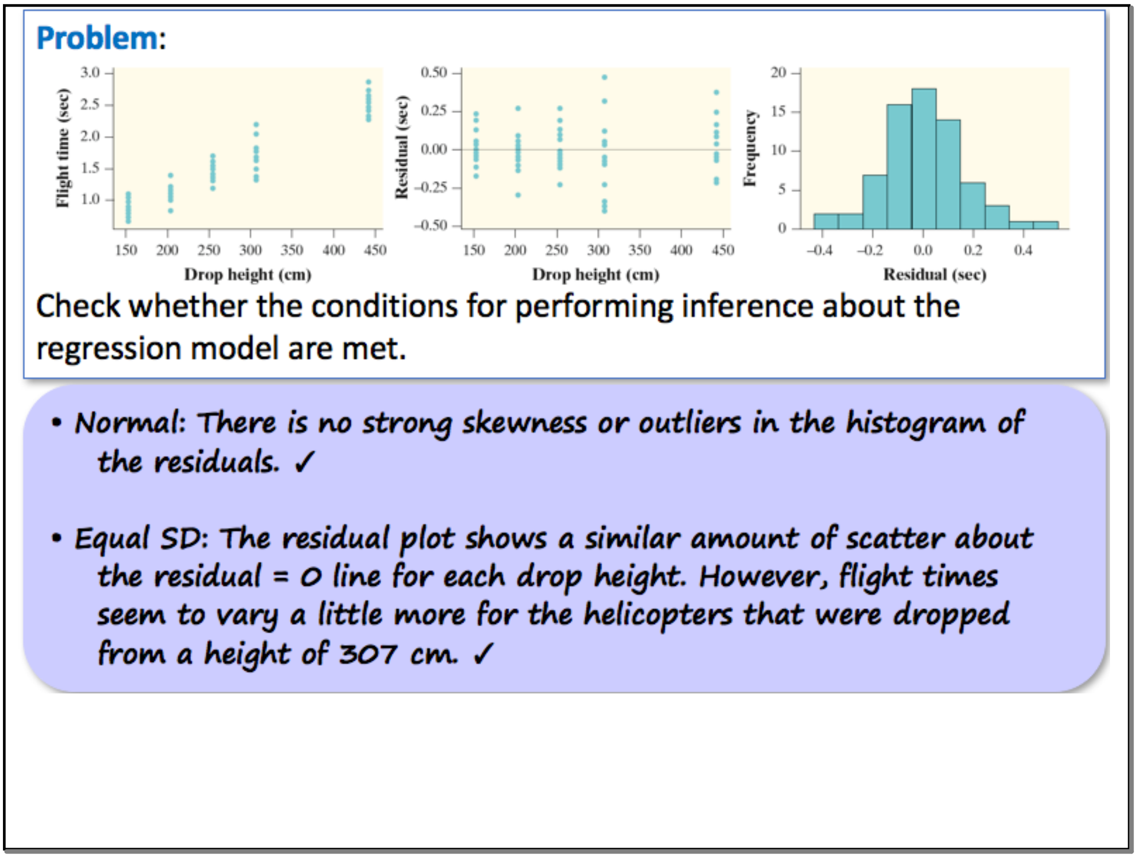
(a) With your group, discuss whether all of the conditions are met for regressions inference

Problem:



Check whether the conditions for performing inference about the regression model are met.

- *Linear: The scatterplot shows a clear linear form and there is no leftover curved pattern in the residual plot.* ✓
- *Independent: Because the helicopters were released in a random order and no helicopter was used twice, knowing the result of one observation should not help us predict the value of another observation.* ✓

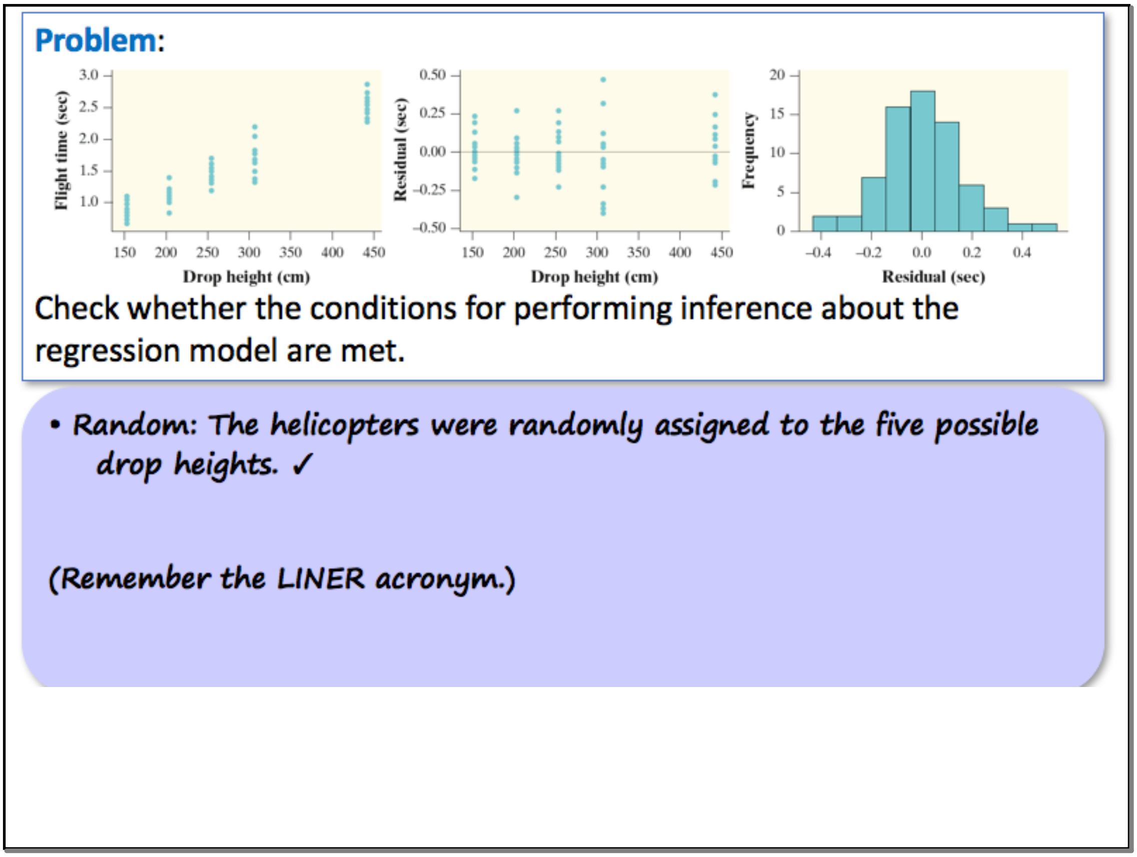


Problem:

Check whether the conditions for performing inference about the regression model are met.

- *Normal: There is no strong skewness or outliers in the histogram of the residuals. ✓*

- *Equal SD: The residual plot shows a similar amount of scatter about the residual = 0 line for each drop height. However, flight times seem to vary a little more for the helicopters that were dropped from a height of 307 cm. ✓*



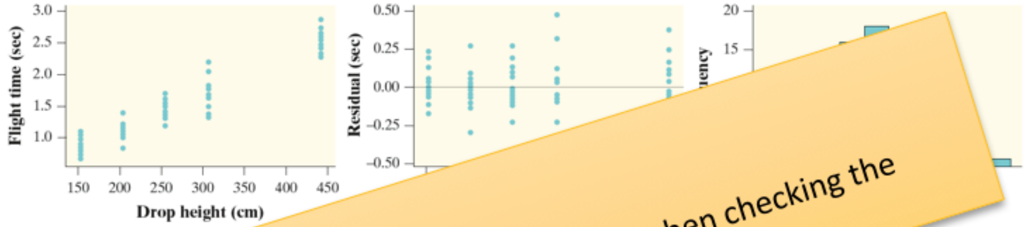
Problem:

Check whether the conditions for performing inference about the regression model are met.

- *Random: The helicopters were randomly assigned to the five possible drop heights. ✓*

(Remember the LINER acronym.)

Problem:



Check whether t… …t the

regres…

CAUTION:

Don't overreact to minor issues in the graphs when checking the Normal and Equal SD conditions

…andomly assigned to the five possible

(Remember the LINER acronym.)