

STAT 216 Coursepack



Spring 2025
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

This resource was developed by Melinda Yager, Jade Schmidt, and Stacey Hancock in 2021 to accompany the online textbook: Hancock, S., Carnegie, N., Meyer, E., Schmidt, J., and Yager, M. (2021). *Montana State Introductory Statistics with R*. Montana State University. <https://mtstateintrostats.github.io/IntroStatTextbook/>.

This resource is released under a Creative Commons BY-NC-SA 4.0 license unless otherwise noted.

Contents

Preface	1
1 Exploratory Data Analysis and Inference for a Quantitative Response with Independent Samples	2
1.1 Vocabulary Review and Key Topics	2
1.2 Video Notes: Inference for Independent Samples	5
1.3 Activity 24: Does behavior impact performance?	14
1.4 Activity 25: Moon Phases and Virtual Reality	19
1.5 Module 12 Lab: Trustworthiness	23
2 Exploratory Data Analysis and Inference for Two Quantitative Variables	28
2.1 Vocabulary Review and Key Topics	28
2.2 Video Notes: Regression and Correlation	31
2.3 Activity 26: Moneyball — Linear Regression	51
2.4 Activity 27: IPEDS (continued)	55
2.5 Activity 28: Prediction of Crocodilian Body Size	62
2.6 Activity 29: Golf Driving Distance	67
2.7 Module 13 Lab: Big Mac Index	73

Preface

This coursepack accompanies the textbook for STAT 216: Montana State Introductory Statistics with R, which can be found at <https://mtstateintrostats.github.io/IntroStatTextbook/>. The syllabus for the course (including the course calendar), data sets, and links to D2L Brightspace, Gradescope, and the MSU RStudio server can be found on the course webpage: <https://math.montana.edu/courses/s216/>. Other notes and review materials are linked in D2L.

Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. In addition to the in-class activities for the course, video notes are provided to aid in taking notes while you complete the required videos. Bring this workbook with you to class each class period, and take notes in the workbook as you would your own notes. A well-written completed workbook will provide an optimal study guide for exams!

All activities and labs in this coursepack will be completed during class time. Parts of each lab will be turned in on Gradescope. To aid in your understanding, read through the introduction for each activity before attending class each day.

STAT 216 is a 3-credit in-person course. In our experience, it takes six to nine hours per week outside of class to achieve a good grade in this class. By “good” we mean at least a C because a grade of D or below does not count toward fulfilling degree requirements. Many of you set your goals higher than just getting a C, and we fully support that. You need roughly nine hours per week to review past activities, read feedback on previous assignments, complete current assignments, and prepare for the next day’s class. A typical week in the life of a STAT 216 student looks like:

- *Prior to class meeting:*
 - Read assigned sections of the textbook, using the provided reading guides to take notes on the material.
 - Watch the provided videos, taking notes in the coursepack.
 - Read through the introduction to the day’s in-class activity.
 - Read through the week’s homework assignment and note any questions you may have on the content.
- *During class meeting:*
 - Work through the guided activity, in-class activity or weekly lab with your classmates and instructor, taking detailed notes on your answers to each question in the activity.
- *After class meeting:*
 - Complete any parts of the activity you did not complete in class.
 - Review the activity solutions in the Math and Stat Center, and take notes on key points.
 - Complete any remaining assigned readings for the week.
 - Complete the week’s homework assignment.

Exploratory Data Analysis and Inference for a Quantitative Response with Independent Samples

1.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a categorical explanatory variable and a quantitative response variable for independent samples.

1.1.1 Key topics

Module 12 will cover exploratory data analysis and both simulation-based and theory-based methods of inference for a quantitative response variable with independent samples. The **summary measure** for a quantitative response with independent samples is a **difference in means**.

- Notation for a difference in sample means: $\bar{x}_1 - \bar{x}_2$, where 1 represents the 1st group of the explanatory variable and 2 represents the 2nd group
- Notation for a difference in population means: $\mu_1 - \mu_2$

Types of plots for a quantitative response variable with independent samples:

- Side-by-side boxplots
- Stacked histograms
- Stacked dotplots

R code to find the summary statistics for a quantitative response variable with independent samples:

```
object %>%  
  reframe(favstats(response ~ explanatory))
```

1.1.2 Vocabulary

Plotting two categorical variables

- **Side-by-side boxplots**: plots a boxplot of the five number summary for each categorical level. R code to create side-by-side boxplots:

```
object %>% # Data set piped into...  
  ggplot(aes(y = response, x = explanatory))+ # Identify variables  
  geom_boxplot()+ # Tell it to make a box plot  
  labs(title = "Don't forget to include a title", # Title: should include the type of plot,  
        # observational units, variables  
        x = "x-axis label", # x-axis label  
        y = "y-axis label") # y-axis label
```

Hypothesis Testing

- **Hypotheses in notation for a paired mean difference:** In the hypotheses below, the **null value** is equal to zero.

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{or} \quad H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 - \mu_2 \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} 0 \quad \text{or} \quad H_A : \mu_1 \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} \mu_2$$

Simulation Hypothesis Testing

- R code for simulation methods to find the p-value using the `two_mean_test` function in the `catstats` package.

```
two_mean_test(response~explanatory, #Enter the names of the variables
  data = object, # Enter the name of the dataset
  first_in_subtraction = "xx", # First outcome in order of subtraction
  number_repetitions = 10000, # Number of simulations
  as_extreme_as = -xx, # Observed statistic
  direction = "xx") # Direction of alternative: "greater", "less", or "two-sided"
```

Simulation Confidence Interval

- R code to find the simulation confidence interval using the `twomean_bootstrap_CI` function from the `catstats` package.

```
two_mean_bootstrap_CI(response ~ explanatory, #Enter the name of the variables
  data = object, # Enter the name of the data set
  first_in_subtraction = "xx", # First value in order of subtraction
  number_repetitions = 10000, # Number of simulations
  confidence_level = xx)
```

- Review how to interpret the confidence interval for two groups from Module 8

Theory-based methods

- **Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution:**
 - **Independence:** The sample's observations are independent, e.g., are from a simple random sample and there is independence between groups. (*Remember:* This also must be true to use simulation methods!)
 - **Normality Condition:** Either the sample observations come from a normally distributed population or we have a large enough sample size. When we have two samples, we need to check this condition for each group! To check this condition, use the following rules of thumb (for both n_1 and n_2):
 - * $n < 30$: The distribution of the sample must be approximately normal with no outliers.
 - * $30 \leq n < 100$: We can relax the condition a little; the distribution of the sample must have no extreme outliers or skewness.
 - * $n \geq 100$: Can assume the sampling distribution of \bar{x} is nearly normal, even if the underlying distribution of individual observations is not.

- **t-distribution:** a theoretical distribution that is symmetric with a given degrees of freedom smallest sample size minus 1 ($n - 1$)

$$t_{n-1}$$

- Calculation of standard error:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Calculation of the standardized difference in sample mean:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)}$$

- The p-value can be found by using the pt function.
 - Enter the value of the standardized statistic for xx
 - Enter the df smallest ($n - 1$) for yy
 - If a greater than alternative, change lower.tail = TRUE to FALSE.
 - If a two-sided test, multiply by 2.

```
pt(xx, df = yy, lower.tail=TRUE)
```

Theory-based methods to find the confidence interval

- Calculation of the confidence interval for a difference in sample means

$$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE(\bar{x}_1 - \bar{x}_2)$$

- R code to find the multiplier for the confidence interval using theory-based methods.
 - qt will give you the multiplier using the t-distribution with smallest $n - 1$ df (enter for yy)
 - Enter the percentile for the given confidence level

```
qt(percentile, df=yy, lower.tail=FALSE)
```

1.2 Video Notes: Inference for Independent Samples

Read Chapters 19 and 20 in the course textbook. Use the following videos to complete the video notes for Module 10.

1.2.1 Course Videos

- 19.1
- 19.2
- 19.3 Theory Tests
- 19.4 Theory Interval

Single categorical, single quantitative variable with independent samples

- In this module, we will study inference for a _____ explanatory variable and a _____ response variable where the two groups are _____.
- Independent groups: When the measurements in one sample are not related to the measurements in the other sample.
- Two random samples taken separately from two populations and the same response variable is recorded. Compare the average number of sick days off from work for people who had a flu shot and people who didn't.
- Participants are randomly assigned to one of two treatment conditions, and the same response variable is recorded.

Rather than analyzing the differences as a single mean we will calculate summary statistics on each sample.

Example: Fifty-one (51) college students volunteered to look at impacts on memorization, specifically if putting letters into recognizable patterns (like FBI, CIA, EDA, CDC, etc.) would increase the number letters memorized. (Miller 1956) The college students were randomly assigned to either a recognizable or non-recognizable letter group. After a period of study time, the number of letters memorized was collected on each study. Is there evidence that putting letters into recognizable letter groups improve memory?

- The summary measure for two independent groups is the _____ in _____.

Notation for Independent Groups

- Population mean for group 1:
- Population mean for group 2:
- Sample mean for group 1:
- Sample mean for group 2:
- Sample difference in means:

- Population standard deviation for group 1:
- Population standard deviation for group 2:
- Sample standard deviation for group 1:
- Sample standard deviation for group 2:
- Sample size for group 1:
- Sample size for group 2:

Why should we treat this as two independent groups rather than paired data?

Hypothesis Testing

Conditions:

- Independence: the response for one observational unit will not influence the outcome for another observational unit

Null hypothesis assumes “no effect”, “no difference”, “nothing interesting happening”, etc.

Always of form: “parameter” = null value

H_0 :

H_A :

- Research question determines the alternative hypothesis.

Write the null and alternative hypotheses for the letters study:

In notation:

H_0 :

H_A :

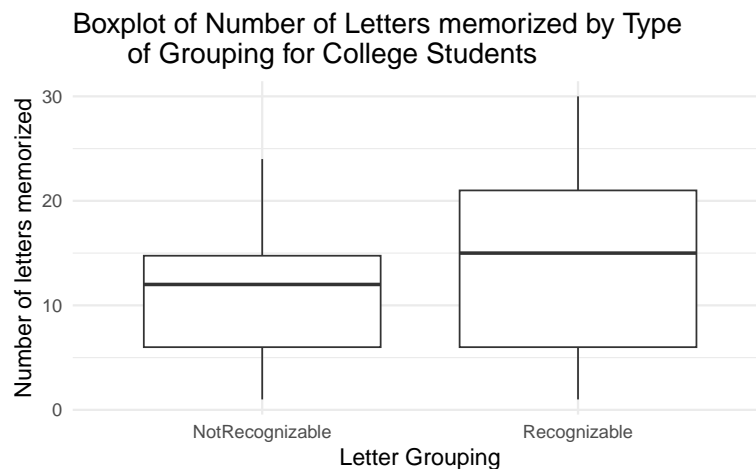
```
letters<-read.csv("data/letters.csv")
letters %>%
  reframe(favstats(Memorized~Grouped))
```

```
#>           Grouped min Q1 median   Q3 max    mean      sd  n missing
#> 1 NotRecognizable   1  6    12 14.75  24 11.15385 6.576883 26      0
#> 2   Recognizable   1  6    15 21.00  30 14.32000 8.518216 25      0
```

Summary statistic:

Interpret the summary statistic in context of the problem:

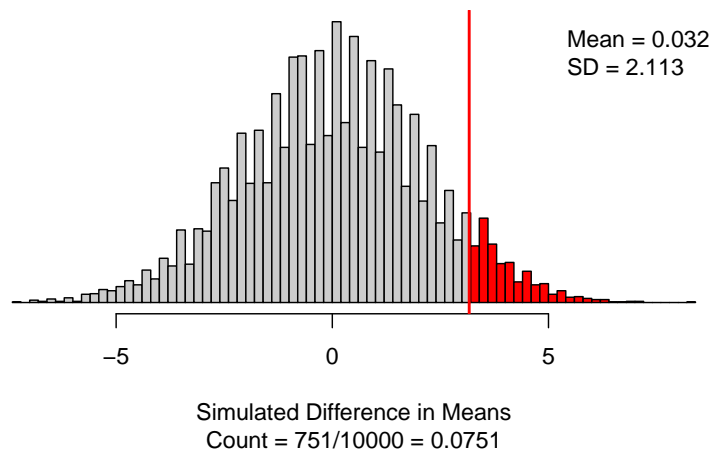
```
letters%>%  
  ggplot(aes(y = Memorized, x = Grouped)) + #Enter the name of the explanatory and response variable  
  geom_boxplot()+  
  labs(title = "Boxplot of Number of Letters memorized by Type  
    of Grouping for College Students", #Title your plot  
    y = "Number of letters memorized", #y-axis label  
    x = "Letter Grouping") #x-axis label
```



Simulation-based method

- Simulate many samples assuming $H_0 : \mu_1 = \mu_2$
 - Write the response variable values on cards
 - Mix the explanatory variable groups together
 - Shuffle cards into two explanatory variable groups to represent the sample size in each group (n_1 and n_2)
 - Calculate and plot the simulated difference in sample means from each simulation
 - Repeat 1000 times (simulations) to create the null distribution
 - Find the proportion of simulations at least as extreme as $\bar{x}_1 - \bar{x}_2$

```
set.seed(216)  
two_mean_test(Memorized~Grouped, #Enter the names of the variables  
  data = letters, # Enter the name of the dataset  
  first_in_subtraction = "Recognizable", # First outcome in order of subtraction  
  number_repetitions = 10000, # Number of simulations  
  as_extreme_as = 3.166, # Observed statistic  
  direction = "greater") # Direction of alternative: "greater", "less", or "two-sided"
```



Explain why the null distribution is centered at the value of zero:

Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Confidence interval

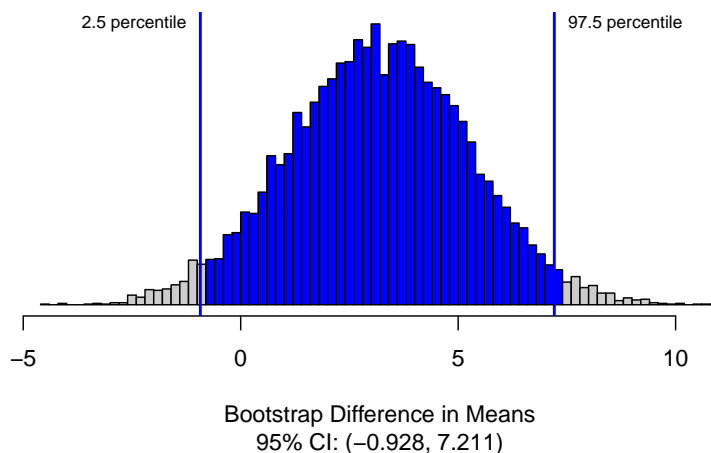
To estimate the difference in true mean we will create a confidence interval.

Simulation-based method - Video 19.2

- Write the response variable values on cards
- Keep explanatory variable groups separate
- Sample with replacement n_1 times in explanatory variable group 1 and n_2 times in explanatory variable group 2
- Calculate and plot the simulated difference in sample means from each simulation
- Repeat 1000 times (simulations) to create the bootstrap distribution
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

For the letters example, we will estimate the difference in true mean number of letters recognized for students given recognizable letter groupings and students given non-recognizable letter groupings.

```
set.seed(216)
two_mean_bootstrap_CI(Memorized ~ Grouped, #Enter the name of the variables
  data = letters, # Enter the name of the data set
  first_in_subtraction = "Recognizable", # First value in order of subtraction
  number_repetitions = 10000, # Number of simulations
  confidence_level = 0.95)
```



Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Theory-based method - Video 19.3 Theory Tests

Example: Every year, orange and black monarch butterflies migrate from their summer breeding grounds in the US and Canada to mountain forests in central Mexico, where they hibernate for the winter. Due to abnormal weather patterns and drought affecting monarch habitats and feeding grounds, the population of monarch butterflies is estimated to have decreased by 53% from the 2018-2019 wintering season to the 2019-2020 wintering season (WWF, 2020). While conservationists often resort to captive-rearing with the goal of raising biologically indistinct individuals for release into the wild, tagging studies have shown that captive-reared monarchs have lower migratory success compared to wild monarchs. For this study, the researchers raised 67 monarchs (descended from wild monarchs) from eggs to maturity and then compared them to a group of 40 wild-caught monarchs. The researchers want to explore whether the maximum grip strength (how many Newtons a butterfly exerts at the moment of release when gently tugged from a mesh-covered perch) differs between captive-reared and wild-caught monarchs. Use Captive – Wild for order of subtraction.

Write the null and alternative hypotheses in notation.

H_0 :

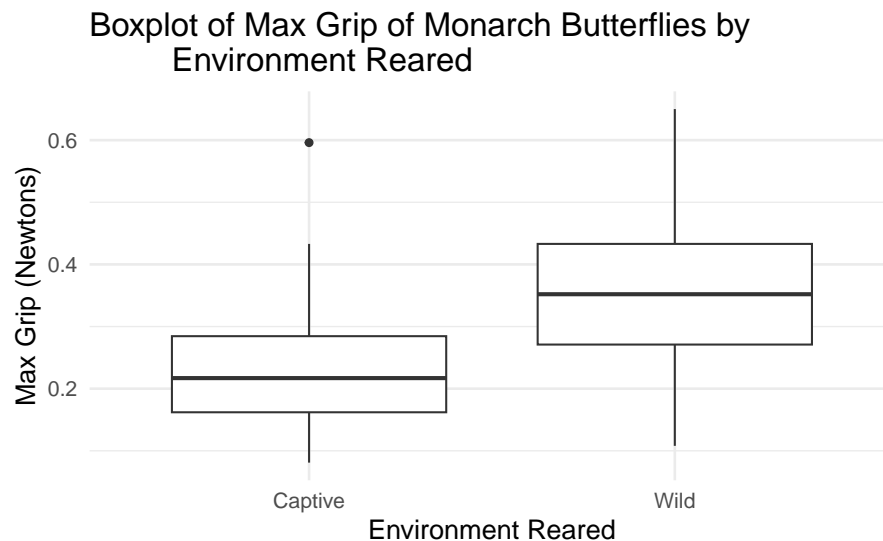
H_A :

```
butterfly <- read.csv("data/butterfly1.csv")

butterflies <- butterfly %>% na.omit() %>%
  rename(Monarch_Group = "Monarch.Group",
         MaxGrip = "Max.Grip.Strength..N.") %>%
  mutate(Monarch_Group = factor(Monarch_Group),
         Sex = factor(Sex)) %>%
  mutate(Monarch_Group = fct_collapse(Monarch_Group, "Captive" = c("Incubator - Fall conditions", "Reari

butterflies %>%
  reframe(favstats(MaxGrip~Monarch_Group))
```

```
#>   Monarch_Group  min    Q1 median    Q3   max    mean    sd  n missing
#> 1      Captive 0.081 0.162  0.217 0.2845 0.596 0.2363731 0.09412948 67      0
#> 2        Wild 0.108 0.271  0.352 0.4330 0.650 0.3607500 0.14066796 40      0
```



Conditions:

- Independence: the response for one observational unit will not influence the outcome for another observational unit
- Large enough sample size

Like with paired data the t-distribution can be used to model the difference in means.

- For independent samples we use the _____ - distribution with _____ degrees of freedom to approximate the sampling distribution.

Theory-based test:

- Calculate the standardized statistic
- Find the area under the t-distribution with the smallest $n - 1$ df $[\min(n_1 - 1, n_2 - 1)]$ at least as extreme as the standardized statistic

Equation for the standard error of the difference in sample mean:

Equation for the standardized difference in sample mean:

Are the conditions met to analyze the butterfly data using theory based-methods?

Calculate the standardized difference in mean max grip strength.

- First calculate the $SE(\bar{x}_1 - \bar{x}_2)$

- Then calculate the T-score

What theoretical distribution should we use to find the p-value?

To find the theory-based p-value:

```
pt(-5, df=39, lower.tail=FALSE)*2
```

```
#> [1] 1.999987
```

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Confidence Interval - Video 19.3 Theory Intervals

- Calculate the interval centered at the sample statistic
statistic \pm margin of error

Using the butterfly data, calculate the 99% confidence interval.

```
butterflies %>%  
  reframe(favstats(MaxGrip~Monarch_Group))
```

```
#>   Monarch_Group  min    Q1 median    Q3   max      mean      sd n missing  
#> 1      Captive 0.081 0.162  0.217 0.2845 0.596 0.2363731 0.09412948 67      0  
#> 2       Wild 0.108 0.271  0.352 0.4330 0.650 0.3607500 0.14066796 40      0
```

- Need the t^* multiplier for a 99% confidence interval from a t-distribution with _____ df.

```
qt(0.995, df=39, lower.tail = TRUE)
```

```
#> [1] 2.707913
```

- We will use the same value for the $SE(\bar{x}_1 - \bar{x}_2)$ as calculated for the standardized statistic.

Calculate the margin of error for a 99% confidence interval for the parameter of interest.

Calculate a 99% confidence interval for the parameter of interest.

1.2.2 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. Why is the recognizable letter study analyzed as two independent groups rather than paired data?
2. Write out the equation for the standard error for a difference in sample means.

1.3 Activity 24: Does behavior impact performance?

1.3.1 Learning outcomes

- Create a side-by-side boxplot of one categorical explanatory variable and one quantitative response variable
- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a simulation-based hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a difference in means.
- Use bootstrapping to find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.
- Use a confidence interval to determine the conclusion of a hypothesis test.

1.3.2 Terminology review

In today's activity, we will use simulation-based methods to analyze the association between one categorical explanatory variable and one quantitative response variable, where the groups formed by the categorical variable are independent. Some terms covered in this activity are:

- Independent groups
- Difference in means

To review these concepts, see Chapter 19 in the textbook.

1.3.3 Behavior and Performance

A study in the Academy of Management Journal (Porath 2017) investigated how rude behaviors influence a victim's task performance. Randomly selected college students enrolled in a management course were randomly assigned to one of two experimental conditions: rudeness condition (45 students) and control group (53 students). Each student was asked to write down as many uses for a brick as possible in five minutes; this value (total number of uses) was used as a performance measure for each student, where higher values indicate better performance. During this time another individual showed up late for class. For those students in the rudeness condition, the facilitator displayed rudeness by berating the students in general for being irresponsible and unprofessional (due to the late-arriving person). No comments were made about the late-arriving person for students in the control group. Is there evidence that the average performance score for students in the rudeness condition is lower than for students in the control group? Use the order of subtraction of rudeness – control.

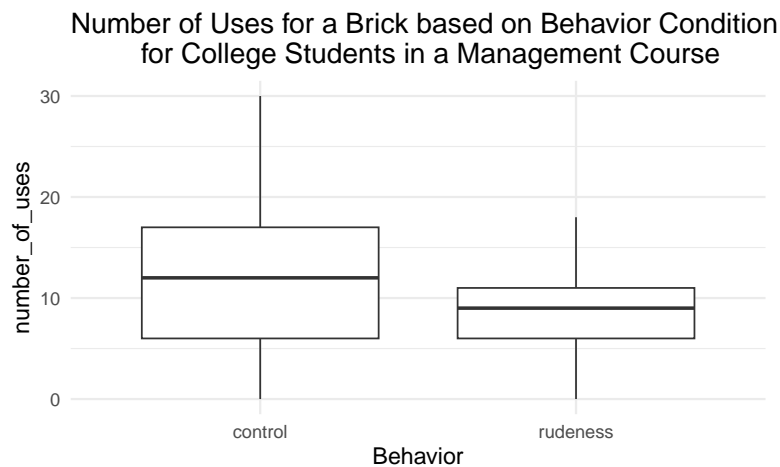
- Download the R script file from D2L and upload to the RStudio server
- Highlight and run lines 1–7 to load the data

```
# Read in data set  
rude <- read.csv("https://math.montana.edu/courses/s216/data/rude.csv")
```

To create a plot of the data and a table of summary statistics.

- Highlight and run lines 11–19

```
# Side-by-side box plots
rude %>%
ggplot(aes(x = condition, y = number_of_uses)) +
  geom_boxplot() +
  labs(title = "Number of Uses for a Brick based on Behavior Condition
for College Students in a Management Course",
x = "Behavior")
# Summary statistics
rude %>%
  reframe(favstats(number_of_uses ~ condition))
#>   condition min  Q1 median Q3 max    mean    sd  n missing
#> 1   control   0   6    12  17  30 11.811321 7.382559 53      0
#> 2  rudeness   0   6     9  11  18  8.511111 3.992164 45      0
```



Quantitative variables review

1. Compare the distributions of the number of bricks between the two treatment conditions.
 - What is the shape of each group?
 - Which group has the higher center?
 - What group has the larger spread?
 - Does either distribution have outliers?
2. Is this an experiment or an observational study? Justify your answer.

3. Explain why this is two independent samples and not paired data.

Ask a research question

In this study we are assessing the difference in true mean number of uses for a brick given by college students enrolled in a management course assigned to a rudeness condition and for those assigned to a control group.

4. What assumption are we making about the difference in true mean?

5. Write the alternative hypothesis in notation.

Numerically Summarize the data

6. Calculate the summary statistic of interest (difference in means). What is the appropriate notation for this statistic?

Interpret this calculated value.

7. Write out the parameter of interest for this study in context of the study.

- To write in context:
 - Population word (true, long-run, population)
 - Summary measure (depends on the type of data)
 - Context
 - * Observational units
 - * Variable(s)

In this study we are assessing the difference in true mean number of uses for a brick given by college students enrolled in a management course assigned to a rudeness condition and for those assigned to a control group.

Use statistical inferential methods to draw inferences from the data

Hypothesis test Remember that the null distribution is created based on the assumption the null hypothesis is true. In this study, the null hypothesis states that there is no association between the two variables. This means that the values observed in the data set would have been the same regardless of the behavior condition.

To demonstrate this simulation, we could create cards to simulate a sample.

- Write the number of uses for a brick given by each student on one card.
- Mix together and shuffle into two piles, one with 45 cards to represent the rudeness condition and one with 53 cards to represent the control group.
- Calculate the difference in mean number of uses for a brick (rudeness - control)

We will use the `two_mean_test()` function in R (in the `catstats` package) to simulate the null distribution of differences in sample means and compute a p-value.

- Fill in the response and explanatory variable names
- Fill in the missing values/names for the xx's in the R script file
- Highlight and run lines 25–30

```
set.seed(216)
two_mean_test(response~explanatory, #Enter the names of the variables
               data = rude, # Enter the name of the dataset
               first_in_subtraction = "xx", # First outcome in order of subtraction
               number_repetitions = 1000, # Number of simulations
               as_extreme_as = xx, # Observed statistic
               direction = "xx") # Direction of alternative: "greater", "less", or "two-sided"
```

8. Report the p-value. Based off of this p-value, write a conclusion to the hypothesis test.

Confidence interval We will use the `two_proportion_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of differences in sample proportions and calculate a confidence interval. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `rude`), the outcome for the explanatory variable that is first in subtraction, number of repetitions, the outcome for the response variable that is a success (the count for the numerator when calculating a sample proportion), and the confidence level as a decimal.

The response variable name is `number_of_uses` and the explanatory variable name is `condition`.

9. What values should be entered for each of the following into the simulation to create a 99% confidence interval?
 - First in subtraction (What is the outcome for the explanatory variable that is used as first in the order of subtraction? "rudeness" or "control"):
 - Number of repetitions:

- Confidence level (entered as a decimal):

Using the R script file for this activity, enter your answers for question 9 in place of the xx's to produce the bootstrap distribution with 1000 simulations; highlight and run lines 35–39.

```
two_mean_bootstrap_CI(response ~ explanatory, #Enter the name of the variables
                      data = rude, # Enter the name of the data set
                      first_in_subtraction = "xx", # First value in order of subtraction
                      number_repetitions = 1000, # Number of simulations
                      confidence_level = xx)
```

10. Where is the bootstrap distribution centered? Explain why.
11. Report the bootstrap 99% confidence interval.
12. What percentile of the bootstrap distribution does the upper value of the confidence interval represent?
13. Interpret the 99% confidence interval.

1.3.4 Take-home messages

1. This activity differs from the activities in Module 11 because the responses are independent, not paired. These data are analyzed as a difference in means, not a mean difference.
2. To create one simulated sample on the null distribution for a difference in sample means, label cards with the response variable values from the original data. Mix cards together and shuffle into two new groups of sizes n_1 and n_2 . Calculate and plot the difference in means.
3. To create one simulated sample on the bootstrap distribution for a difference in sample means, label $n_1 + n_2$ cards with the original response values. Keep groups separate and randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Calculate and plot the resampled difference in means.

1.3.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered

1.4 Activity 25: Moon Phases and Virtual Reality

1.4.1 Learning outcomes

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a theory-based hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a difference in means.
- Use theory-based methods to find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.
- Use a confidence interval to determine the conclusion of a hypothesis test.

1.4.2 Terminology review

In today's activity, we will use theory-based methods to analyze the association between one categorical explanatory variable and one quantitative response variable, where the groups formed by the categorical variable are independent. Some terms covered in this activity are:

- Difference in means
- Independence within and between groups
- Normality

To review these concepts, see Chapter 19 in the textbook.

1.4.3 Moon Phases and Virtual Reality

In a study comparing immersive virtual reality (VR) to traditional hands-on methods, researchers recruited 115 undergraduate students to assess the effectiveness of these approaches in teaching complex scientific concepts like Moon phases (Madden 2020). Participants were randomly assigned to experience either a VR simulation replicating the Sun-Earth-Moon system or a hands-on activity where they physically manipulated models to observe Moon phases. The students were given a 14 multiple choice question quiz about Moon phases and the Moon's motion relative to the Earth to evaluate their understanding of Moon phases and the Moon's motion. Each question had only one correct answer, and the participant's score was the sum of the number of correct answers, with all questions weighted equally (with a maximum score of 14). Is there evidence of a difference, on average, in student learning comparing those using VR methods to those using the traditional method? Use order of subtraction VR – Hands-on.

- Download the RScript file and dataset from D2L and upload to the RStudio server
- Open the RScript file
- Enter the name of the data set for `datasetname` in line 8
- Highlight and run lines 1–8

```
moon <- read.csv("datasetname.csv")
```

1. Write out the parameter of interest in words in context of the study.

2. Write out the null hypothesis in notation for this study. Be sure to clearly identify the subscripts.

3. Write out the alternative hypothesis in words for this study.

The sampling distribution for $\bar{x}_1 - \bar{x}_2$ can be modeled using a normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution:

- **Independence:** The sample's observations are independent
- **Normality:** Each sample should be approximately normal or have a large sample size. For *each* sample:
 - $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.
 - $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers, then we typically assume the sampling distribution of \bar{x} is nearly normal, even if the underlying distribution of individual observations is not.
 - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of \bar{x} is nearly normal, even if the underlying distribution of individual observations is not.

To create the plots of the data:

- Enter a title in line 15 for the plot between the quotations
- Highlight and run lines 12 - 17

```
moon %>% # Data set piped into...
  ggplot(aes(y = TestScore, x = Method))+ # Identify variables
  geom_boxplot()+ # Tell it to make a box plot
  labs(title = "Don't forget to title your plot!", # Title
       x = "Methods", # x-axis label
       y = "Test Score (points)") # y-axis label
```

To find the summary statistic:

- Enter the response and explanatory variable names in line 22
- Highlight and run lines 21–22

```
moon %>%
  reframe(favstats(response~explanatory))
```

4. Can theory-based methods be used to analyze these data?

5. Calculate the summary statistic (difference in means) for this study. Use appropriate notation with clearly defined subscripts.

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the difference in means we will calculate:

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)},$$

where the standard error of the difference in means is calculated using:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

6. Calculate the standard error for the difference in sample means.

7. Calculate the standardized statistic for the difference in sample means.

To find the degrees of freedom to use for the t-distribution, we need to use the group with the smallest sample size and subtract 1. (**df** = minimum of $n_1 - 1$ or $n_2 - 1$).

- Enter the value of the standardized statistic for **xx**
- Enter the df for **yy**
- Highlight and run line 27

```
2*pt(xx, df=yy, lower.tail=FALSE)
```

8. Report the p-value for the study. Why did we multiply by two to find the p-value?

To calculate a theory-based 95% confidence interval for a difference in means, use the formula:

$$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE(\bar{x}_1 - \bar{x}_2)$$

We will need to find the t^* multiplier using the function **qt()**.

- Enter the percentile to find the multiplier for a 95% confidence level

- Enter the degrees of freedom for yy
- Highlight and run line 32

```
qt(percentile, df = yy, lower.tail=TRUE)
```

9. Calculate the margin of error for a 95% confidence interval.
10. Calculate the 95% confidence interval.
11. Write a conclusion to the test.

1.4.4 Take-home messages

1. In order to use theory-based methods for independent groups, the normality condition must be met for each sample.
2. A T-score is compared to a t -distribution with the minimum $n - 1$ df in order to calculate a one-sided p-value. To find a two-sided p-value using theory-based methods we need to multiply the one-sided p-value by 2.
3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t -distribution with the minimum $n - 1$ df.

1.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered

1.5 Module 12 Lab: Trustworthiness

1.5.1 Learning outcomes

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a theory-based hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a difference in means.
- Use theory-based methods to find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.
- Use a confidence interval to determine the conclusion of a hypothesis test.

1.5.2 Trustworthiness

Researchers in India wanted to find out how trustworthy famous YouTubers are (Kalra 2022). They went through a process in which they collected data on many videos from famous YouTubers to determine a trustworthiness score. Scientists randomly selected videos from famous YouTubers (>1000 subscribers) to include in the study. There were many different factors that went into calculating the trustworthiness score. Researchers also recorded if YouTubers were a subject matter expert (SME) or not a subject matter expert (non-SME). An example of an SME would be if one of your statistics professors made a YouTube video of how to do hypothesis testing. An example of someone who isn't an SME would be if one of your friends who has never taken a civil engineering class in their life decided to make a YouTube video about how to build a bridge. There were 621 Youtubers who are SMEs in the sample and 1026 who aren't SMEs. Is there evidence of a difference in mean trustworthiness score between subject matter experts (SME) YouTubers and non-SME YouTubers? Use SME – Non -SME as the order of subtraction

1. **Write out the parameter of interest in words in context of the study.**
2. Write out the null hypothesis in notation for this study. Be sure to clearly identify the subscripts.
3. Write out the alternative hypothesis in words for this study.

The sampling distribution for $\bar{x}_1 - \bar{x}_2$ can be modeled using a normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}_1 - \bar{x}_2$ to follow an approximate normal distribution:

- **Independence:** The sample's observations are independent
- **Normality:** Each sample should be approximately normal or have a large sample size. For *each* sample:

- $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.
 - $30 \leq n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers, then we typically assume the sampling distribution of \bar{x} is nearly normal, even if the underlying distribution of individual observations is not.
 - $n \geq 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of \bar{x} is nearly normal, even if the underlying distribution of individual observations is not.
- Upload and open the R script file for Module 10 lab. Upload the csv file, `Trustworthiness.csv`.
 - Enter the name of the data set for `datasetname` in the R script file in line 10.
 - Write a title for the boxplots in line 14.
 - Highlight and run lines 1–16 to load the data and create plots of the data.

```
trust <- read.csv("datasetname")
trust %>% # Data set piped into...
  ggplot(aes(y = Trustworthiness_Video, x = Creator_SME))+ # Identify variables
  geom_boxplot()+ # Tell it to make a box plot
  labs(title = "Don't forget to include a title", # Title: should include the type of plot,
        # observational units, variables
        x = "Whether the Creator is SME", # x-axis label
        y = "Trustworthiness Score") # y-axis label
```

4. Is the independence condition met? Explain your answer.
5. Check that the normality condition is met to use theory-based methods to analyze these data.

- Enter the name of the explanatory variable for `explanatory` and the name of the response variable for `response` in line 22.
- Highlight and run lines 21–22 to get the summary statistics for the data.

```
trust %>%
  reframe(favstats(response~explanatory))
```

6. Calculate the summary measure (difference in means) for this study. Use appropriate notation with clearly defined subscripts.

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the difference in means we will calculate:

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)},$$

where the standard error of the difference in means is calculated using:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

7. Calculate the standard error for the difference in sample means.

8. Calculate the standardized statistic for the difference in sample means.

9. When we are comparing two quantitative variables to find the degrees of freedom to use for the t-distribution, we need to use the group with the smallest sample size and subtract 1. (**df** = minimum of $n_1 - 1$ or $n_2 - 1$). Calculate the **df** for this study.

10. Using the provided R script file, enter the T-score (for **xx**) and the **df** calculated in question 9 for **yy** into the **pt()** function to find the p-value. Highlight and run line 27. Report the p-value calculated.

```
2*pt(xx, df=yy, lower.tail=FALSE)
```

11. Explain why we multiplied by 2 in the code above.

12. Do you expect the 95% confidence interval to contain the null value of zero? Explain your answer.

To calculate a theory-based 95% confidence interval for a difference in means, use the formula:

$$(\bar{x}_1 - \bar{x}_2) \pm (t^* \times SE(\bar{x}_1 - \bar{x}_2))$$

We will need to find the t^* multiplier using the function **qt()**. For a 95% confidence level, we are finding the t^* value at the 97.5th percentile with (**df** = minimum of $n_1 - 1$ or $n_2 - 1$).

- Enter the appropriate percentile value (as a decimal) for **xx** and degrees of freedom for **yy** into the **qt()** function at line 32 to find the appropriate t^* multiplier

```
qt(xx, df = yy, lower.tail=FALSE)
```

13. Report the t^* multiplier for the 95% confidence interval.

14. Calculate the 95% confidence interval using theory-based methods.
15. Do the results of the CI agree with the p-value? Explain your answer.
16. What type of error may be possible?
17. Write a paragraph summarizing the results of the study as if you are reporting the results to your supervisor.
Upload a copy of your paragraph to Gradescope for your group. Be sure to describe:
- Summary statistic and interpretation
 - P-value and interpretation
 - Statement about probability or proportion of samples
 - Statistic (summary measure and value)
 - Direction of the alternative
 - Null hypothesis (in context)
 - Confidence interval and interpretation
 - How confident you are (e.g., 90%, 95%, 98%, 99%)
 - Parameter of interest
 - Calculated interval
 - Order of subtraction when comparing two groups
 - Conclusion (written to answer the research question)
 - Amount of evidence
 - Parameter of interest
 - Direction of the alternative hypothesis
 - Scope of inference

Paragraph continued:

Exploratory Data Analysis and Inference for Two Quantitative Variables

2.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of two quantitative variables.

2.1.1 Key topics

Module 13 will cover exploratory data analysis and both simulation-based and theory-based methods of inference for two quantitative variables.

Types of plot for two quantitative variables

- **Scatterplot:** plots (x,y) pairs of observations
- Four characteristics of scatterplots
 - Form (linear or non-linear)
 - Direction (positive or negative)
 - Strength (weak, moderate, or strong)
 - Outliers?

R code to create a scatterplot:

```
object %>% # Pipe data set into...
ggplot(aes(x = explanatory, y = response))+ # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "x-axis label", # Label x-axis
       y = "y-axis label", # Label y-axis
       title = "Don't forget to add a title!") +
  # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line
```

Summary measures:

- **Slope of the regression line:** measures the magnitude and direction of the linear relationship between two quantitative variables

$$\widehat{response} = b_0 + b_1 \times explanatory$$

- Parameter notation for slope: β_1
- Sample notation for slope: b_1
- R code to create the linear model

```
linearmodel <- lm(response~explanatory, data=object)
round(summary(linearmodel)$coefficients,3) # Display coefficient
```

- **Correlation:** measures the strength and direction of the linear relationship between two quantitative variables
 - Parameter notation: ρ
 - Sample notation: r
- **Coefficient of determination:** measures the percent of total variability in the response variable that is explained by the relationship with the explanatory variable

$$r^2 = (r)^2 = \frac{SST - SSE}{SST} = \frac{s_y^2 - s_{residual}^2}{s_y^2}$$

Simulation Hypothesis Testing

- Can test either slope or correlation - both test for a linear relationship between two quantitative variables

Hypotheses for slope:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \left\{ \begin{array}{l} < \\ \neq \\ < \end{array} \right\} 0$$

Hypotheses for correlation:

$$H_0 : \rho = 0$$

$$H_A : \rho \left\{ \begin{array}{l} < \\ \neq \\ < \end{array} \right\} 0$$

- R code for simulation methods to find the p-value using the `regression_test` function in the `catstats` package.

```
regression_test(response~explanatory, # response ~ explanatory
  data = object, # Name of data set
  direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
  summary_measure = "xx", # "slope" or "correlation"
  as_extreme_as = xx, # Observed slope or correlation
  number_repetitions = 10000) # Number of simulated samples for null distribution
```

Conditions necessary to use theory-based methods

When performing inference on a least squares line, the follow conditions are generally required:

- *Independent observations* (for both simulation-based and theory-based methods): individual data points must be independent.
 - Check this assumption by investigating the sampling method and determining if the observational units are related in any way.
- *Linearity* (for both simulation-based and theory-based methods): the data should follow a linear trend.

- Check this assumption by examining the scatterplot of the two variables, and a scatterplot of the residuals (on the y -axis) versus the fitted values (on the x -axis). The pattern in the residual plot should display a horizontal line.
- *Constant variability* (for theory-based methods only): the variability of points around the least squares line remains roughly constant
 - Check this assumption by examining a scatterplot of the residuals (on the y -axis) versus the fitted values (on the x -axis). The variability in the residuals around zero should be approximately the same for all fitted values.
- *Nearly normal residuals* (for theory-based methods only: residuals must be nearly normal).
 - Check this assumption by examining a histogram of the residuals, which should appear approximately normal.

Theory-based Methods to find the p-value

- To find the value of the standardized statistic to test the slope we will use,

$$T = \frac{\text{slope estimate} - \text{nullvalue}}{SE} = \frac{b_1 - 0}{SE(b_1)}.$$

* Use the standard error estimate of the slope from the linear model output

- The p-value can be found from the linear model output or by using the `pt` function.
 - Enter the value of the standardized statistic for `xx`
 - Enter the df ($n - 2$) for `yy`

```
pt(xx, df = yy, lower.tail=TRUE)
```

Simulation methods to find the confidence interval

- R code to find the simulation confidence interval using the `regression_bootstrap_CI` function from the `catstats` package.

```
regression_bootstrap_CI(response~explanatory, # response ~ explanatory
  data = object, # Name of data set
  confidence_level = xx, # Confidence level as decimal
  summary_measure = "xx", # Slope or correlation
  number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```

Theory-based methods to find the confidence interval

- Calculation of the confidence interval for slope

$$b_1 \pm t^* \times SE(b_1)$$

- R code to find the multiplier for the confidence interval using theory-based methods.
 - `qt` will give you the multiplier using the t-distribution with $n-2$ df (enter for `yy`)
 - Enter the percentile for the given confidence level

```
qt(percentile, df=yy, lower.tail=FALSE)
```

2.2 Video Notes: Regression and Correlation

Read Chapters 6, 7, and 8 in the course textbook. Use the following videos to complete the video notes for Module 13.

2.2.1 Course Videos

- 6.1
- 6.2
- 6.3
- Ch 7

Summary measures and plots for two quantitative variables - Videos 6.1 - 6.3

Example: Data were collected from 1236 births between 1960 and 1967 in the San Francisco East Bay area to better understand what variables contributed to child birthweight, as children with low birthweight often suffer from an array of complications later in life (“Child Health and Development Studies,” n.d.). There were some missing values in the study and with those observations removed we have a total of 1223 births.

```
babies<-read.csv("data/babies.csv") %>%
  drop_na(bwt) %>%
  drop_na(gestation)
glimpse(babies)
#> Rows: 1,223
#> Columns: 8
#> $ case      <int> 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
#> $ bwt       <int> 120, 113, 128, 108, 136, 138, 132, 120, 143, 140, 144, 141, ~
#> $ gestation <int> 284, 282, 279, 282, 286, 244, 245, 289, 299, 351, 282, 279, ~
#> $ parity    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
#> $ age       <int> 27, 33, 28, 23, 25, 33, 23, 25, 30, 27, 32, 23, 36, 30, 38, ~
#> $ height    <int> 62, 64, 64, 67, 62, 62, 65, 62, 66, 68, 64, 63, 61, 63, 63, ~
#> $ weight    <int> 100, 135, 115, 125, 93, 178, 140, 125, 136, 120, 124, 128, 9~
#> $ smoke     <int> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, ~
```

Here you see a glimpse of the data. The 1223 rows correspond to the sample size. The case variable is labeling each pregnancy 1 through 1223. Then 7 variables are recorded. birthweight (bwt), length of gestation in days, parity is called an indicator variable telling us if the pregnancy was a first pregnancy (labeled as 0) or not (labeled as 1) were recorded about the child and pregnancy. The age, height, and weight were recorded for the mother giving birth, as was smoke, another indicator variable where 0 means the mother did not smoke during pregnancy, and 1 indicates that she did smoke while pregnant.

Type of plot

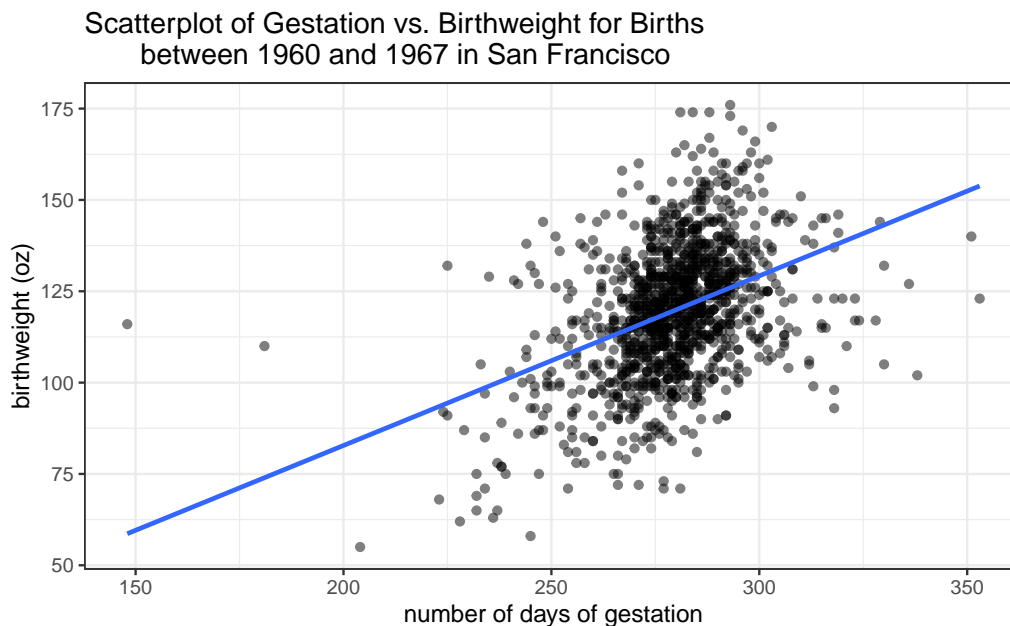
A _____ is used to display the relationship between two _____ variables.

Four characteristics of the scatterplot:

- Form:
- Direction:
- Strength:
- Outliers:
 - Influential points: outliers that change the regression line; far from the line of regression
 - High leverage points: outliers that are extreme in the x- axis; far from the mean of the x-axis

The following shows a scatterplot of length of gestation as a predictor of birthweight.

```
babies %>% # Data set pipes into...
ggplot(aes(x = gestation, y = bwt))+ # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "number of days of gestation", # Label x-axis
       y = "birthweight (oz)", # Label y-axis
       title = "Scatterplot of Gestation vs. Birthweight for Births
               between 1960 and 1967 in San Francisco") +
  # Be sure to title your plots with the type of plot, observational units, variable(s)
  geom_smooth(method = "lm", se = FALSE) + # Add regression line
  theme_bw()
```



Describe the scatterplot using the four characteristics of a scatterplot.

The summary measures for two quantitative variables are:

- _____
- _____
- _____

Notation:

- Population slope:
- Population correlation:
- Sample slope:
- Sample correlation:

Correlation

Correlation is always between the values of _____ and _____.

- Measures the _____ and _____ of the linear relationship between two quantitative variables.
- The stronger the relationship between the variables the closer the value of _____ is to _____ or _____.
- The sign gives the _____.

The following code creates a correlation matrix between different quantitative variables in the data set.

```
babies %>%  
  select(c("gestation", "age", "height", "weight", "bwt")) %>%  
  cor(use="pairwise.complete.obs") %>%  
  round(3)
```

```
#>      gestation    age height weight    bwt  
#> gestation      1.000 -0.056  0.064  0.022 0.408  
#> age           -0.056  1.000 -0.005  0.147 0.029  
#> height         0.064 -0.005  1.000  0.436 0.201  
#> weight         0.022  0.147  0.436  1.000 0.154  
#> bwt            0.408  0.029  0.201  0.154 1.000
```

The value of correlation between gestation and birthweight is _____. This shows a _____, _____ relationship between gestation and birthweight.

Slope

- Least-squares regression line: $\hat{y} = b_0 + b_1 \times x$ (put y and x in the context of the problem) or $\widehat{response} = b_0 + b_1 \times \text{explanatory}$
- \hat{y} or $\widehat{response}$ is
- b_0 is
- b_1 is
- x or explanatory is

- The estimates for the linear model output will give the value of the _____ and the _____.
- Interpretation of slope: an increase in the _____ variable of 1 unit is associated with an increase/decrease in the _____ variable by the value of slope, on average.
- Interpretation of the y-intercept: for a value of 0 for the _____ variable, the predicted value for the _____ variable would be the value of y-intercept.
- We can predict values of the _____ variable by plugging in a given _____ variable value using the least squares equation line.
- A prediction of a response variable value for an explanatory value outside the range of x values is called _____.
- To find how far the predicted value deviates from the actual value we find the _____.

- To find the least squares regression line the line with the _____ SSE is found.

SSE = sum of squared errors

- To find SSE, the residual for each data point is found, squared and all the squared residuals are summed together

The linear model output for this study is given below:

```
# Fit linear model: y ~ x
babiesLM <- lm(bwt ~ gestation, data=babies)
round(summary(babiesLM)$coefficients,3) # Display coefficient summary
```

```
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  -10.064      8.322  -1.209   0.227
#> gestation      0.464      0.030  15.609   0.000
```

Write the least squares equation of the line.

Interpret the slope in context of the problem.

Interpret the y-intercept in context of the problem.

Predict the birthweight for a birth with a baby born at 310 days gestation.

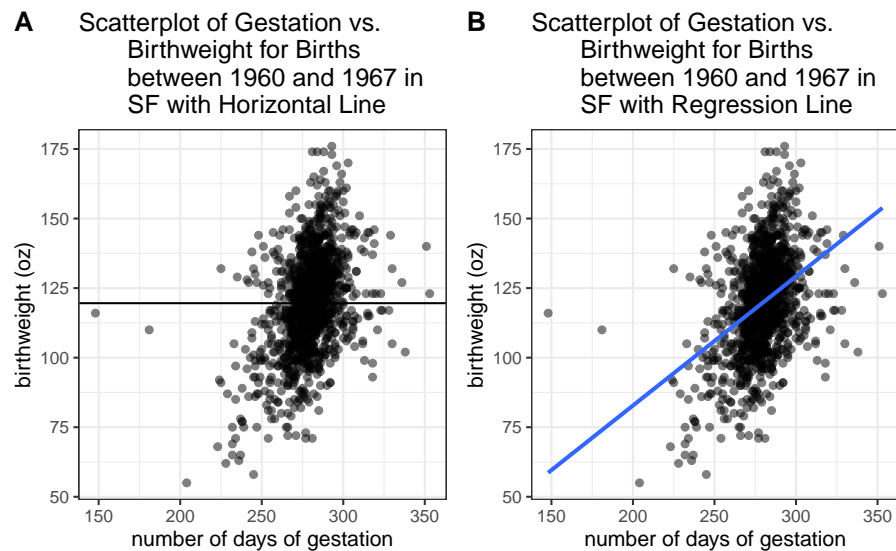
Calculate the residual for a birth of a baby with a birthweight of 151 ounces and born at 310 days gestation.

Is this value (310, 151) above or below the line of regression? Did the line of regression overestimate or underestimate the birthweight?

Coefficient of Determination

The coefficient of determination can be found by squaring the value of correlation, using the variances for each variable or using the SSE (sum of squares error) and SST (sum of squares total)

- $r^2 = (r)^2 = \frac{SST - SSE}{SST} = \frac{s_y^2 - s_{residual}^2}{s_y^2}$
- The coefficient of determination measures the _____ of total variation in the _____ variable that is explained by the changes in the _____ variable.



The value for SST was calculated as 406753.48. The value for SSE was calculated as 339092.13.

Calculate the coefficient of determination between gestation and birthweight.

Interpret the coefficient of determination between gestation and birthweight.

Multivariable plots - Video Chapter7

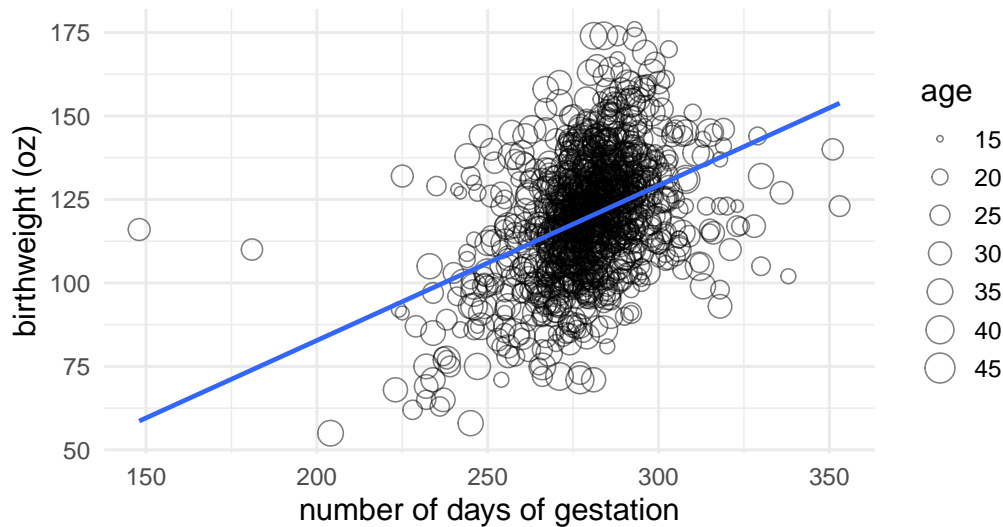
Aesthetics: visual property of the objects in your plot

- Position on the axes: groups for _____ variables, or a number line if the variable is _____
- Color or shape - to represent _____ variables
- Size - to represent _____ variables

Adding the quantitative variable maternal age to the scatterplot between gestation and birthweight.

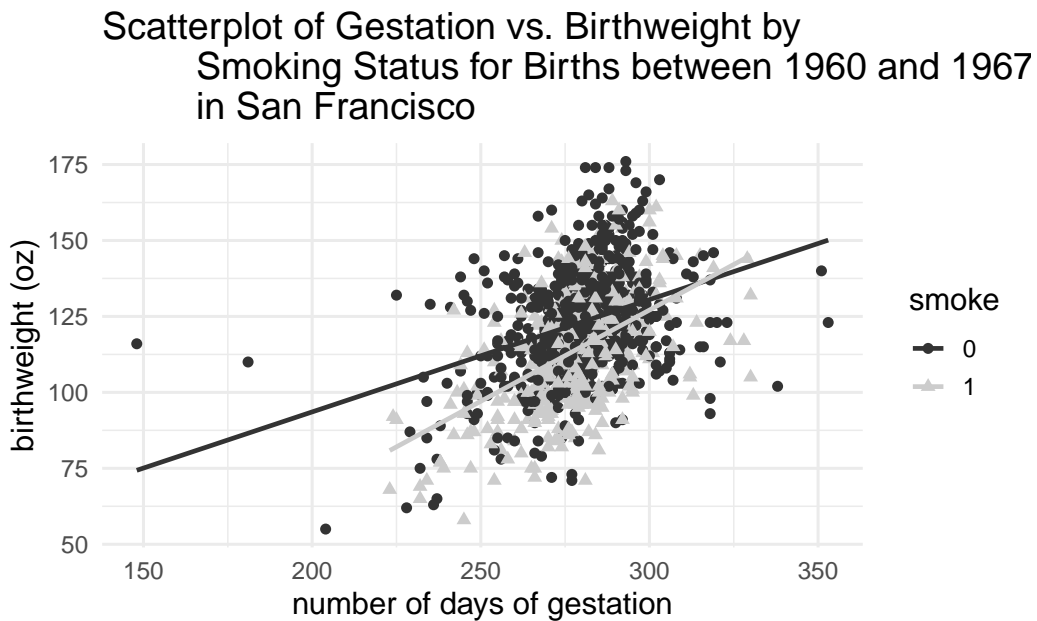
```
babies %>% # Data set pipes into...
ggplot(aes(x = gestation, y = bwt))+ # Specify variables
  geom_point(alpha=0.5, shape=1, aes(size=age)) + # Add scatterplot of points
  labs(x = "number of days of gestation", # Label x-axis
       y = "birthweight (oz)", # Label y-axis
       title = "Scatterplot of Gestation vs. Birthweight by Age
               for Births between 1960 and 1967 in San Francisco") +
  # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line
```

Scatterplot of Gestation vs. Birthweight by Age
for Births between 1960 and 1967 in San Francisco



Let's add the categorical variable, whether a mother smoked, to the scatterplot between gestation and birthweight.

```
babies <- babies %>%  
  mutate(smoke = factor(smoke)) %>%  
  na.omit()  
  
babies %>% # Data set pipes into...  
  ggplot(aes(x = gestation, y = bwt, color = smoke)) + #Specify variables  
  geom_point(aes(shape = smoke), size = 2) + #Add scatterplot of points  
  labs(x = "number of days of gestation", #Label x-axis  
       y = "birthweight (oz)", #Label y-axis  
       title = "Scatterplot of Gestation vs. Birthweight by  
               Smoking Status for Births between 1960 and 1967  
               in San Francisco") +  
  #Be sure to title your plots  
  geom_smooth(method = "lm", se = FALSE) + #Add regression line  
  scale_color_grey()
```



Does the relationship between length of gestation and birthweight appear to depend upon maternal smoking status?

Is the variable smoking status a potential confounding variable?

Adding a categorical predictor:

- Look at the regression line for each level of the _____
- If the slopes are _____, the two predictor variables do not _____ to help explain the response
- If the slopes _____, there is an interaction between the categorical predictor and the relationship between the two quantitative variables.

2.2.2 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. What are the three summary measures for two quantitative variables?
2. What are the four characteristics used to describe a scatterplot?
3. When we add a categorical predictor variable to a scatterplot of two quantitative variables, what summary measure will we compare across the categories to assess the change in the relationship between the two quantitative variables.

2.2.3 Video Notes: Inference for Two Quantitative Variables

Read Chapters 21 and 22 in the course textbook. Use the following videos to complete the video notes for Module 11.

2.2.4 Course Videos

- 21.1
- 21.3
- 21.4TheoryTests
- 21.4TheoryIntervals

Summary measures and plots for two quantitative variables.

Scatterplot:

- Form: linear or non-linear?
- Direction: positive or negative?
- Strength: how clear is the pattern between the two variables?
- Outliers: points that are far from the pattern or bulk of the data
 - Influential points: outliers that are extreme in the x- variable.

The summary measures for two quantitative variables are:

- _____, interpreted as the on average change in the response variable for a one unit increase in the explanatory variable.
- _____, which measures the strength and direction of the linear relationship between two quantitative variables.
- _____, interpreted as the percent of variability in the response variable that is explained by the relationship with the explanatory variable.
- Least-squares regression line: $\hat{y} = b_0 + b_1 \times x$ (put y and x in the context of the problem)

Notation:

- Population slope:
- Population correlation:
- Sample slope:
- Sample correlation:

Example: Oceanic temperature is important for sea life. The California Cooperative Oceanic Fisheries Investigations has measured several variables on the Pacific Ocean for more than 70 years hoping to better understand weather patterns and impacts on ocean life. (“Ocean Temperature and Salinity Study,” n.d.) For this example, we will look at the most recent 100 measurements of salt water salinity (measured in PSUs or

practical salinity units) and the temperature of the ocean measured in degrees Celsius. Is there evidence that water temperature in the Pacific Ocean tends to decrease with higher levels of salinity?

Hypothesis Testing

Null hypothesis assumes “no effect”, “no difference”, “nothing interesting happening”, etc.

Always of form: “parameter” = null value

H_0 :

H_A :

- Research question determines the alternative hypothesis.

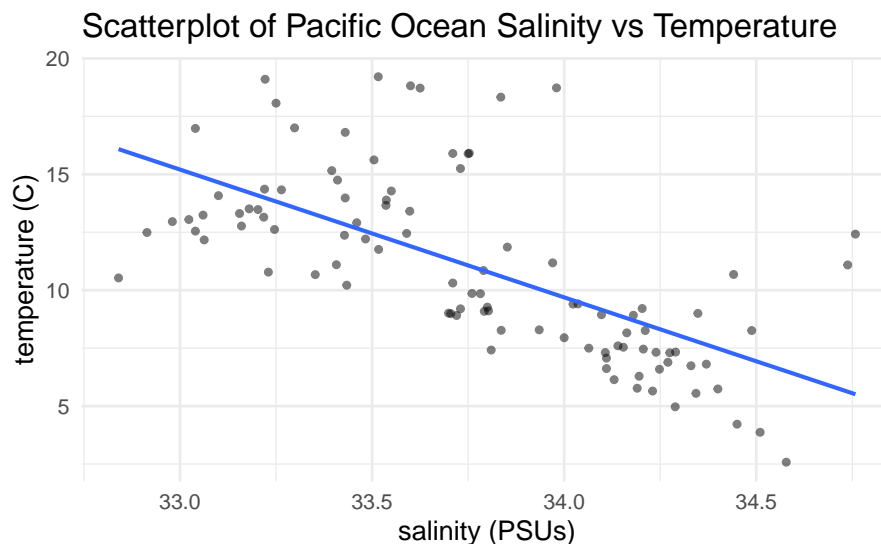
Write the null and alternative for the ocean study:

In notation:

H_0 :

H_A :

```
water %>% # Pipe data set into...
ggplot(aes(x = Salnty, y = T_degC))+ # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "salinity (PSUs)", # Label x-axis
       y = "temperature (C)", # Label y-axis
       title = "Scatterplot of Pacific Ocean Salinity vs Temperature") +
  # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line
```



Describe the four characteristics of the scatterplot:

Linear model output:

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients, 3)
```

```
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  197.156     21.478    9.18      0
#> Salnty       -5.514      0.636   -8.67      0
```

Correlation:

```
cor(T_degC~Salnty, data=water)
```

```
#> [1] -0.6588365
```

Write the least squares equation of the line in context of the problem:

Interpret the value of slope in the context of the problem:

Report and describe the correlation value:

Calculate and interpret the coefficient of determination:

Simulation-based method

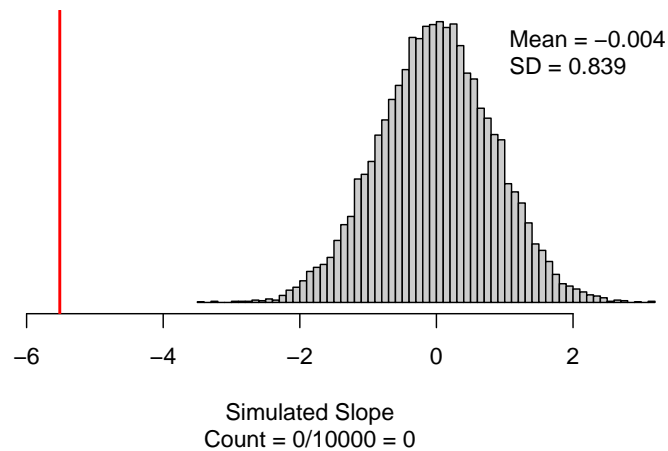
Conditions:

- Independence: the response for one observational unit will not influence another observational unit
- Linear relationship:
- Simulate many samples assuming $H_0 : \beta_1 = 0$ or $H_0 : \rho = 0$
 - Write the response variable values on cards
 - Hold the explanatory variable values constant

- Shuffle a new response variable to an explanatory variable
- Plot the shuffled data points to find the least squares line of regression
- Calculate and plot the simulated slope or correlation from each simulation
- Repeat 1000 times (simulations) to create the null distribution
- Find the proportion of simulations at least as extreme as b_1 or r

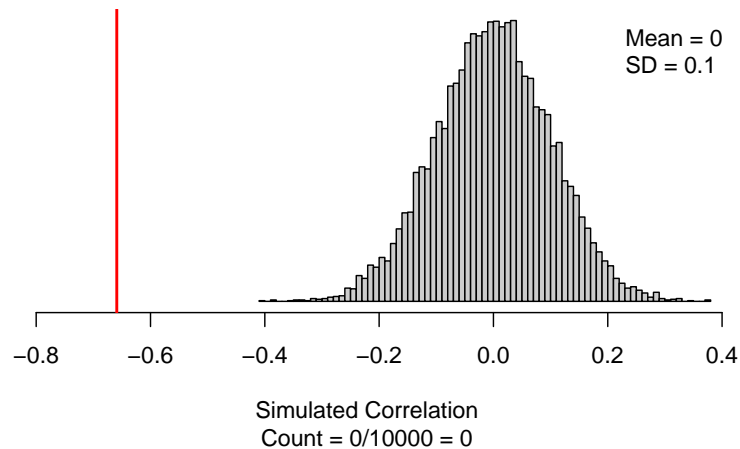
To test slope:

```
set.seed(216)
regression_test(T_degC ~ Salnty, # response ~ explanatory
  data = water, # Name of data set
  direction = "less", # Sign in alternative ("greater", "less", "two-sided")
  summary_measure = "slope", # "slope" or "correlation"
  as_extreme_as = -5.514, # Observed slope or correlation
  number_repetitions = 10000) # Number of simulated samples for null distribution
```



To test correlation:

```
set.seed(216)
regression_test(T_degC~Salnty, # response ~ explanatory
  data = water, # Name of data set
  direction = "less", # Sign in alternative ("greater", "less", "two-sided")
  summary_measure = "correlation", # "slope" or "correlation"
  as_extreme_as = -0.659, # Observed slope or correlation
  number_repetitions = 10000) # Number of simulated samples for null distribution
```



Explain why the null distribution is centered at the value of zero:

Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Confidence interval - Video 21.3

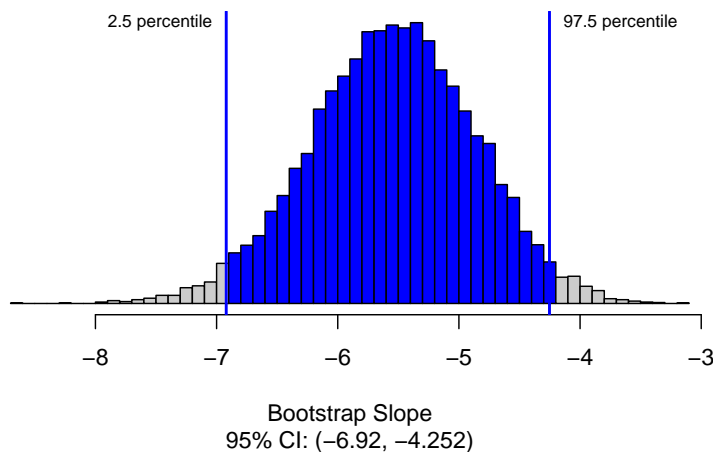
To estimate the true slope (or true correlation) we will create a confidence interval.

Simulation-based method

- Write the explanatory and response value pairs on cards
- Sample pairs with replacement n times
- Plot the resampled data points to find the least squares line of regression
- Calculate and plot the simulated slope (or correlation) from each simulation
- Repeat 1000 times (simulations) to create the bootstrap distribution
- Find the cut-offs for the middle X% (confidence level) in a bootstrap distribution.

Returning to the ocean example, we will estimate the true slope between salinity and temperature of the Pacific Ocean.

```
set.seed(216)
regression_bootstrap_CI(T_degC~Salnty, # response ~ explanatory
  data = water, # Name of data set
  confidence_level = 0.95, # Confidence level as decimal
  summary_measure = "slope", # Slope or correlation
  number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```

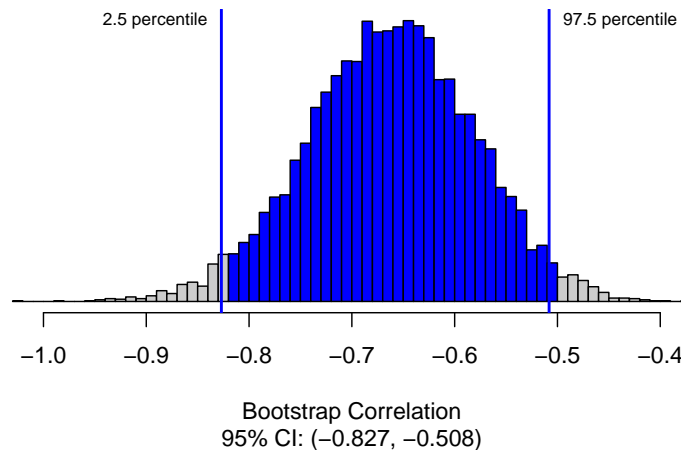


Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Now we will estimate the true correlation between salinity and temperature of the Pacific Ocean.

```
set.seed(216)
regression_bootstrap_CI(T_degC~Salnty, # response ~ explanatory
  data = water, # Name of data set
  confidence_level = 0.95, # Confidence level as decimal
  summary_measure = "correlation", # Slope or correlation
  number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```



Confidence interval interpretation:

- How confident you are (e.g., 90%, 95%, 98%, 99%)
- Parameter of interest
- Calculated interval
- Order of subtraction when comparing two groups

Theory-based method - Video 21.4to21.5TheoryTests

Conditions:

- Linearity (for both simulation-based and theory-based methods): the data should follow a linear trend.
 - Check this assumption by examining the _____ of the two variables, and _____. The pattern in the residual plot should display a horizontal line.

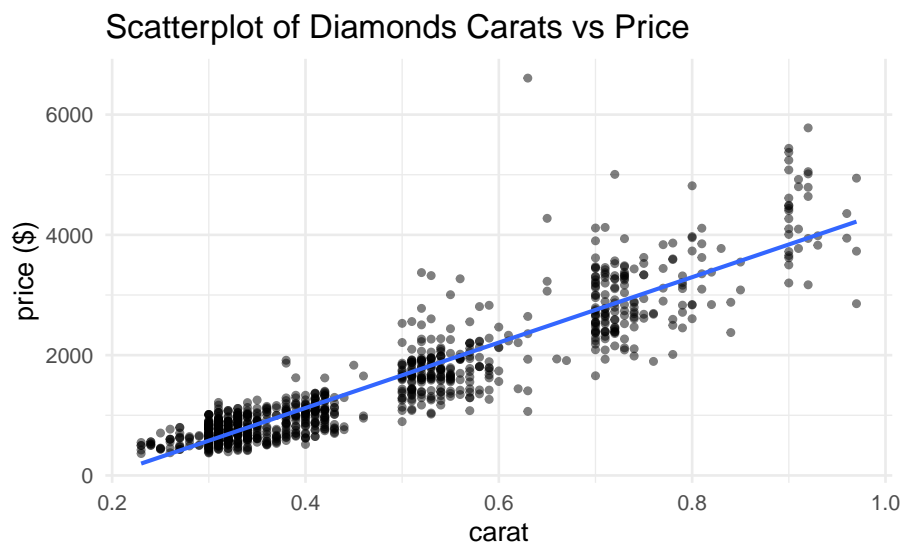
- Independence (for both simulation-based and theory-based methods)
 - One _____ for an observational unit has no impact on _____.
- Constant variability (for theory-based methods only): the variability of points around the least squares line remains roughly constant
 - Check this assumption by examining the _____. The variability in the residuals around zero should be approximately the same for all fitted values.
- Nearly normal residuals (for theory-based methods only): residuals must be nearly normal
 - Check this assumption by examining a _____, which should appear approximately normal

Example:

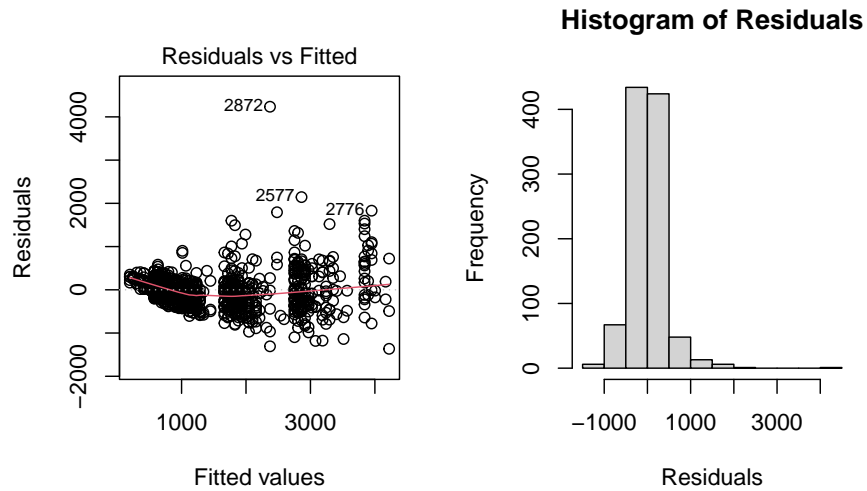
It is a generally accepted fact that the more carats a diamond has, the more expensive that diamond will be. The question is, how much more expensive? Data on thousands of diamonds were collected for this data set. We will only look at one type of cut (“Ideal”) and diamonds less than 1 carat. Does the association between carat size and price have a linear relationship for these types of diamonds? What can we state about the association between carat size and price?

Scatterplot:

```
Diamonds %>% # Pipe data set into...
  ggplot(aes(x = carat, y = price)) + # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "carat", # Label x-axis
       y = "price ($)", # Label y-axis
       title = "Scatterplot of Diamonds Carats vs Price") +
  # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line
```



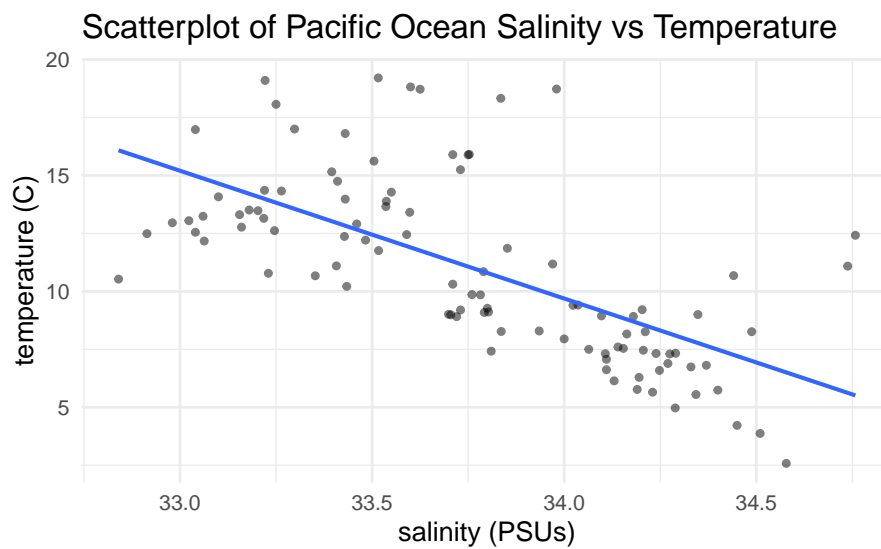
Diagnostic plots:



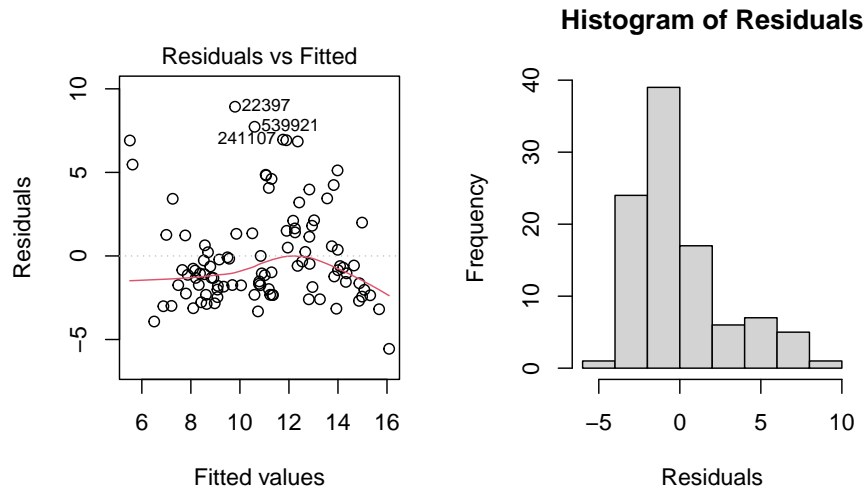
Check the conditions for the ocean data:

Scatterplot:

```
water %>% # Pipe data set into...
ggplot(aes(x = Salnty, y = T_degC)) + # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "salinity (PSUs)", # Label x-axis
       y = "temperature (C)", # Label y-axis
       title = "Scatterplot of Pacific Ocean Salinity vs Temperature") +
  # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line
```



Diagnostic plots:



Like with paired data the t -distribution can be used to model slope and correlation.

- For two quantitative variables we use the _____-distribution with _____ degrees of freedom to approximate the sampling distribution.

Theory-based test:

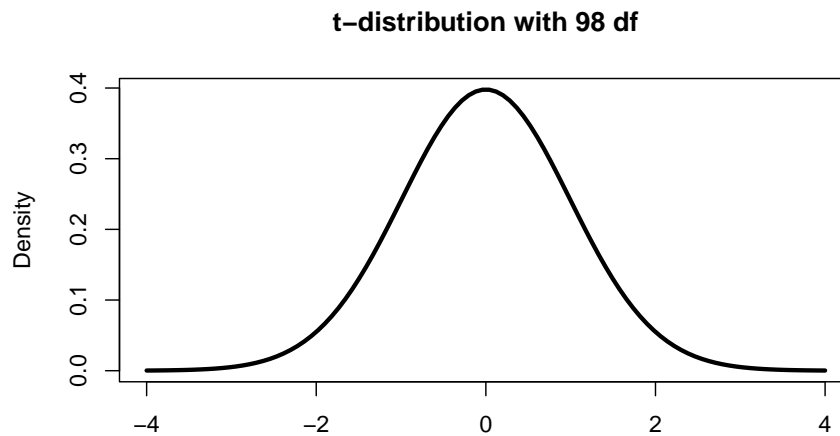
- Calculate the standardized statistic
- Find the area under the t -distribution with $n - 2$ df at least as extreme as the standardized statistic

Equation for the standardized slope:

Calculate the standardized slope for the ocean data

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients,3)
```

```
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  197.156     21.478    9.18      0
#> Salnty       -5.514      0.636   -8.67      0
```



Interpret the standardized statistic:

To find the theory-based p-value:

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response~explanatory)
round(summary(lm.water)$coefficients,3)
```

```
#>           Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  197.156     21.478    9.18     0
#> Salnty       -5.514      0.636   -8.67     0
```

or

```
pt(-8.670, df = 98, lower.tail=TRUE)
#> [1] 4.623445e-14
```

Theory-based method

- Calculate the interval centered at the sample statistic
statistic \pm margin of error

```
lm.water <- lm(T_degC~Salnty, data=water) # lm(response-explanatory)
round(summary(lm.water)$coefficients, 3)
```

```
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   197.156      21.478    9.18     0
#> Salnty        -5.514       0.636   -8.67     0
```

Using the ocean data, calculate a 95% confidence interval for the true slope.

- Need the t^* multiplier for a 95% confidence interval from a t-distribution with _____ df.

```
qt(0.975, df=98, lower.tail = TRUE)
```

```
#> [1] 1.984467
```

2.2.5 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. Explain why theory-based methods should not be used to analyze the salinity study?
2. What is the proper notation for the population slope? Population correlation?

2.3 Activity 26: Moneyball — Linear Regression

2.3.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.
- Use scatterplots to assess the relationship between two quantitative variables.
- Find the estimated line of regression using summary statistics and R linear model (`lm()`) output.
- Interpret the slope coefficient in context of the problem.

2.3.2 Terminology review

In today's activity, we will review summary measures and plots for two quantitative variables. Some terms covered in this activity are:

- Scatterplot
- Least-squares line of regression
- Slope and y -intercept
- Residuals

To review these concepts, see Chapter 6 & 7 in the textbook.

2.3.3 Moneyball

The goal of a Major League baseball team is to make the playoffs. In 2002, the manager of the Oakland A's, Billy Bean, with the help of Paul DePodesta began to use statistics to determine which players to choose for their season. Based on past data, DePodesta determined that to make it to the playoffs, the A's would need to win at least 95 games in the regular season. In order to win more games, they would need to score more runs than they allowed. The Oakland A's won 20 consecutive games and a total of 103 games for the season. The success of this use of sports analytics was portrayed by the 2011 movie, Moneyball. In this study, we will see if there is evidence of a positive linear relationship between the difference in the number of runs scored minus the number of runs allowed (RD) and the number of wins for Major League baseball teams in the years before 2002. Some of the variables collected in the data set baseball consist of the following:

Variable	Description
RA	Runs allowed
RS	Runs scored
OBP	On-base percentage
SLG	Slugging percentage
BA	Batting average
OOBP	Opponent's on-base percentage
OSLG	Opponent's slugging percentage
W	Number of wins in the season
RD	Difference of runs scored minus runs allowed

```
moneyball <- read.csv("data/baseball.csv") # Reads in data set
moneyball$RD <- moneyball$RS - moneyball$RA
moneyball <-
  moneyball %>% # Pipe data set into
  subset(Year < 2002) # Select only years before 2002
```

Vocabulary review

- Use the provided R script file to create a scatterplot to examine the relationship between the difference in number of runs scored minus number of runs allowed and the number of wins by filling in the variable names (RD and W) for explanatory and response in line 14. Note, we are using the difference in runs scores minus runs allowed to predict the number of season wins.
- Highlight and run lines 1–20.

```
moneyball %>% # Data set pipes into...
  ggplot(aes(x = explanatory, y = response))+ # Specify variables
  geom_point() + # Add scatterplot of points
  labs(x = "Difference in number of runs", # Label x-axis
       y = "Number of Season wins", # Label y-axis
       title = "Scatterplot of Run Difference vs. Number of Season Wins for MLB Teams") +
  # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line
```

1. Assess the four features of the scatterplot that describe this relationship.
 - Form (linear, non-linear)
 - Direction (positive, negative)
 - Strength
 - Unusual observations or outliers
2. Based on the plot, does there appear to be an association between run difference and number of season wins? Explain your answer.

Slope

The linear model function in R (`lm()`) gives us the summary for the least squares regression line. The estimate for (**Intercept**) is the y -intercept for the line of least squares, and the estimate for **budget_mil** (the x -variable name) is the value of b_1 , the slope.

- Run lines 24–25 in the R script file to reproduce the linear model output found in the coursepack.

```
# Fit linear model: y ~ x
moneyballLM <- lm(W~RD, data=moneyball)
round(summary(moneyballLM)$coefficients, 3) # Display coefficient summary
```

```
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   80.881      0.131 616.675      0
#> RD            0.106      0.001  81.554      0
```

3. Write out the least squares regression line using the summary statistics provided above in context of the problem.

4. Interpret the value of slope in context of the problem.
5. Using the least squares line from question 3, predict the number of season wins for a MLB team that has a run difference of -66 runs.
6. Predict the number of season wins for a MLB team that has a run difference of 400 runs.
7. The prediction in question 6 is an example of what?

Residuals

The model we are using assumes the relationship between the two variables follows a straight line. The residuals are the errors, or the variability in the response that hasn't been modeled by the regression line.

$$\Rightarrow \text{Residual} = \text{actual y value} - \text{predicted y value}$$

$$e = y - \hat{y}$$

8. The MLB team *Florida Marlins* had a run difference of -66 runs and 79 wins for the season. Find the residual for this MLB team.
9. Did the line of regression overestimate or underestimate the number of wins for the season for this team?

Correlation

The following output shows a correlation matrix between several pairs of quantitative variables.

- Highlight and run lines 29–33 to produce the same table as below.

```
moneyball %>% # Data set pipes into
  select(c("RD", "BA",
           "SLG", "W")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

```
#>      RD    BA   SLG    W
#> RD  1.000 0.442 0.428 0.939
#> BA  0.442 1.000 0.814 0.416
#> SLG 0.428 0.814 1.000 0.406
#> W   0.939 0.416 0.406 1.000
```

10. Report the value of correlation between the run difference and the number of season wins.
11. Calculate the coefficient of determination between the run difference and the number of season wins.
12. Interpret the value of coefficient of determination in context of the study.

2.3.4 Take-home messages

1. Two quantitative variables are graphically displayed in a scatterplot. The explanatory variable is on the x -axis and the response variable is on the y -axis. When describing the relationship between two quantitative variables we look at the form (linear or non-linear), direction (positive or negative), strength, and for the presence of outliers.
2. There are three summary statistics used to summarize the relationship between two quantitative variables: correlation (r), slope of the regression line (b_1), and the coefficient of determination (r^2).

2.3.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

2.4 Activity 27: IPEDS (continued)

2.4.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set with two quantitative variables.
- Find the estimated line of regression using summary statistics and R linear model (`lm()`) output.
- Interpret the slope coefficient in context of the problem.
- Calculate and interpret r^2 , the coefficient of determination, in context of the problem.
- Find the correlation coefficient from R output or from r^2 and the sign of the slope.

2.4.2 Terminology review

In today's activity, we will review summary measures and plots for two quantitative variables. Some terms covered in this activity are:

- Least-squares line of regression
- Slope and y -intercept
- Residuals
- Correlation (r)
- Coefficient of determination (r -squared)

To review these concepts, see Chapter 6 in the textbook.

2.4.3 The Integrated Postsecondary Education Data System (IPEDS)

We will continue to assess the IPEDS data set collected on a subset of institutions that met the following selection criteria (Education Statistics 2018):

- Degree granting
- United States only
- Title IV participating
- Not for profit
- 2-year or 4-year or above
- Has full-time first-time undergraduates

Some of the variables collected and their descriptions are below. Note that several variables have missing values for some institutions (denoted by "NA").

Variable	Description
UnitID	Unique institution identifier
Name	Institution name
State	State abbreviation
Sector	whether public or private
LandGrant	Is this a land-grant institution (Yes/No)
Size	Institution size category based on total student enrolled for credit, Fall 2018: Under 1,000, 1,000 - 4,999, 5,000 - 9,999, 10,000 - 19,999, 20,000 and above
Cost_OutofState	Cost of attendance for full-time out-of-state undergraduate students

Variable	Description
Cost_InState	Cost of attendance for full-time in-state undergraduate students
Retention	Retention rate is the percent of the undergraduate students that re-enroll in the next year
Graduation_Rate	6-year graduation rate for undergraduate students
SATMath_75	75th percentile Math SAT score
ACT_75	75th percentile ACT score

The code below reads in the needed data set, IPEDS_2018.csv, and filters out the 2-year institutions.

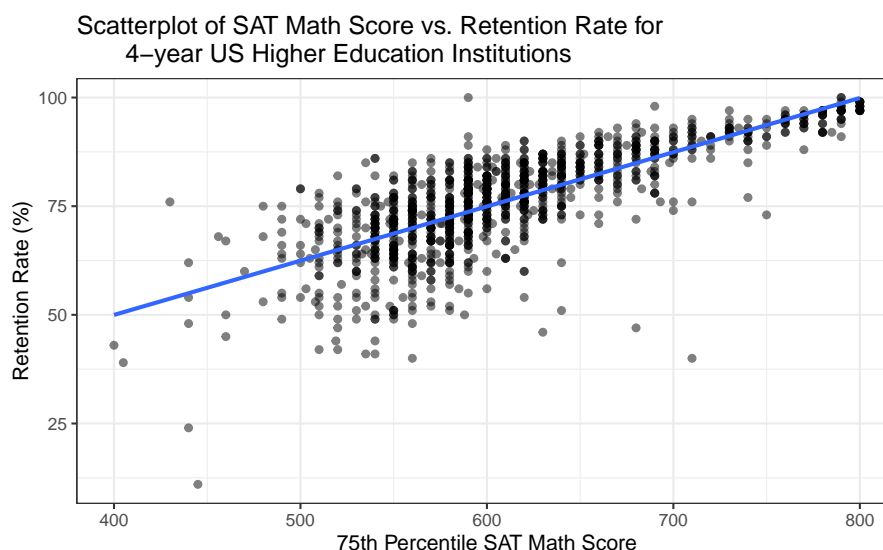
- Highlight and run lines 1 – 11 to load the data set and filter out the 2-year institutions.

```
IPEDS <- read.csv("https://www.math.montana.edu/courses/s216/data/IPEDS_2018.csv")
IPEDS <- IPEDS %>%
  filter(Sector != "Public 2-year") #Filters the data set to remove Public 2-year
IPEDS <- IPEDS %>%
  filter(Sector != "Private 2-year") #Filters the data set to remove Private 2-year
IPEDS <- na.omit(IPEDS)
```

To create a scatterplot of the 75th percentile Math SAT score by retention rate for 4-year US Higher Education Institutions...

- Enter the variable SATMath_75 for explanatory and Retention for response in line 16.
- Highlight and run lines 15–21.

```
IPEDS %>% # Data set pipes into...
  ggplot(aes(x = SATMath_75, y = Retention)) + # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "75th Percentile SAT Math Score", # Label x-axis
       y = "Retention Rate (%)", # Label y-axis
       title = "Scatterplot of SAT Math Score vs. Retention Rate for
               4-year US Higher Education Institutions") +
  # Be sure to title your plots with the type of plot, observational units, variable(s)
  geom_smooth(method = "lm", se = FALSE) + # Add regression line
  theme_bw()
```



1. Describe the relationship between 75th percentile SAT Math score and retention rate.

Slope of the Least Squares Linear Regression Line

There are three summary measures calculated from two quantitative variables: slope, correlation, and the coefficient of determination. We will first assess the slope of the least squares regression line between 75th percentile SAT Math score and retention rate.

- Enter **Retention** for response and **SATMath_75** for explanatory in line 25
- Highlight and run lines 25 – 26 to fit the linear model.

```
# Fit linear model: y ~ x
IPEDSLM <- lm(Retention~SATMath_75, data=IPEDS)
round(summary(IPEDSLM)$coefficients,3) # Display coefficient summary
```

```
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    0.059      1.898    0.031   0.975
#> SATMath_75     0.125      0.003   40.485   0.000
```

2. Write out the least squares regression line using the summary statistics from the R output in context of the problem.

Slope Interpretation: An increase of one point in SAT Math 75th percentile score is associated with an increase in retention rate, on average, of 0.125 percentage points for 4-year higher education institutions.

3. Predict the retention rate for a 4-year US higher education institution with a 75th percentile SAT Math score of 440.
4. Calculate the residual for a 4-year US higher education institution with a 75th percentile SAT Math score of 440 and a retention rate of 24%.

Correlation

Correlation measures the strength and the direction of the linear relationship between two quantitative variables. The closer the value of correlation to +1 or -1, the stronger the linear relationship. Values close to zero indicate a very weak linear relationship between the two variables.

The following output creates a correlation matrix between several pairs of quantitative variables.

```
IPEDS %>% # Data set pipes into
  select(c("Retention", "Cost_InState",
           "Graduation_Rate", "Salary",
           "SATMath_75", "ACT_75")) %>%
```

```
cor(use="pairwise.complete.obs") %>%
round(3)
```

```
#>           Retention Cost_InState Graduation_Rate Salary SATMath_75 ACT_75
#> Retention           1.000         0.388           0.832  0.698         0.767  0.768
#> Cost_InState        0.388         1.000           0.563  0.365         0.502  0.514
#> Graduation_Rate     0.832         0.563           1.000  0.683         0.817  0.833
#> Salary              0.698         0.365           0.683  1.000         0.747  0.706
#> SATMath_75          0.767         0.502           0.817  0.747         1.000  0.920
#> ACT_75              0.768         0.514           0.833  0.706         0.920  1.000
```

5. What is the value of correlation between SATMath_75 and Retention?

Coefficient of determination (squared correlation)

Another summary measure used to explain the linear relationship between two quantitative variables is the coefficient of determination (r^2). The coefficient of determination, r^2 , can also be used to describe the strength of the linear relationship between two quantitative variables. The value of r^2 (a value between 0 and 1) represents the **proportion of variation in the response that is explained by the least squares line with the explanatory variable**. There are two ways to calculate the coefficient of determination:

Square the correlation coefficient: $r^2 = (r)^2$

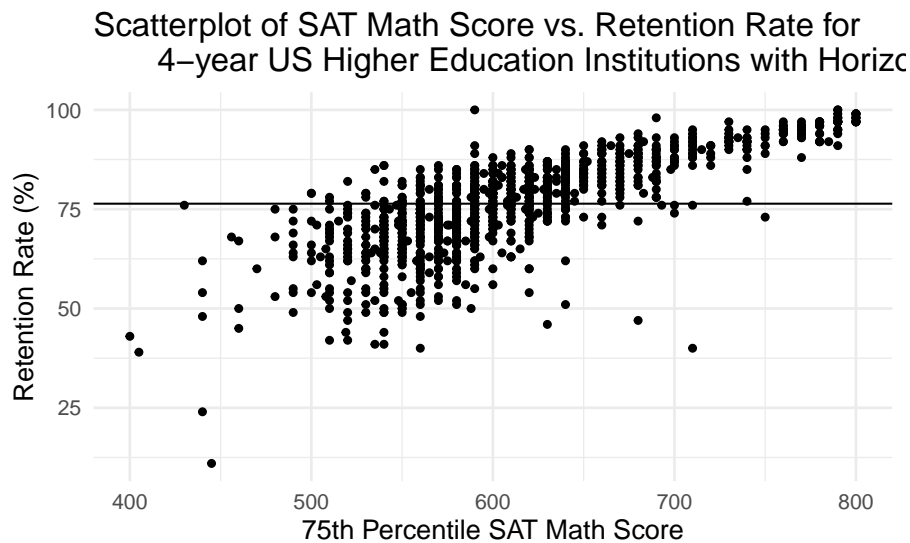
Use the variances of the response and the residuals: $r^2 = \frac{s_y^2 - s_{RES}^2}{s_y^2} = \frac{SST - SSE}{SST}$

6. Use the correlation, r , found in question 5, to calculate the coefficient of determination between SATMath_75 and Retention, r^2 .

7. The variance of the response variable, Retention in \$MM, is $s_{Retention}^2 = 138.386 \%^2$ and the variability in the residuals is $s_{RES}^2 = 56.934 \%^2$. Use these values to calculate the coefficient of determination.

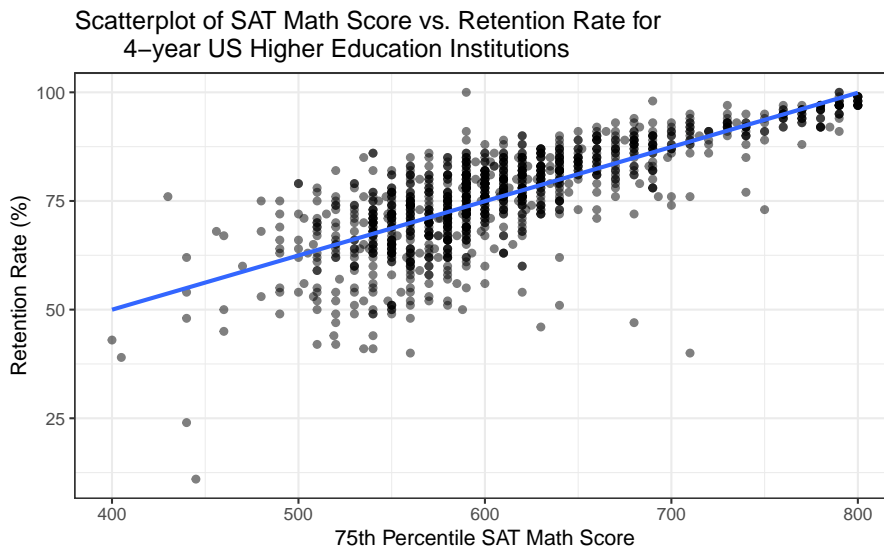
In the next part of the activity we will explore what the coefficient of determination measures.

In the first scatterplot, we see the data plotted with a horizontal line. Note that the regression line in this plot has a slope of zero; this assumes there is no relationship between SATMath_75 and Retention. The value of the y-intercept, 76.387, is the mean of the response variable when there is no relationship between the two variables. To find the sum of squares total (SST) we find the residual ($residual = y - \hat{y}$) for each response value from the horizontal line (from the value of 76.387). Each residual is squared and the sum of the squared values is calculated. The SST gives the **total variability in the response variable, Retention**.



The calculated value for the SST is 158451.8.

This next scatterplot, shows the plotted data with the best fit regression line. This is the line of best fit between budget and revenue and has the smallest sum of squares error (SSE). The SSE is calculated by finding the residual from each response value to the regression line. Each residual is squared and the sum of the squared values is calculated.



The calculated value for the SSE is 65133.022.

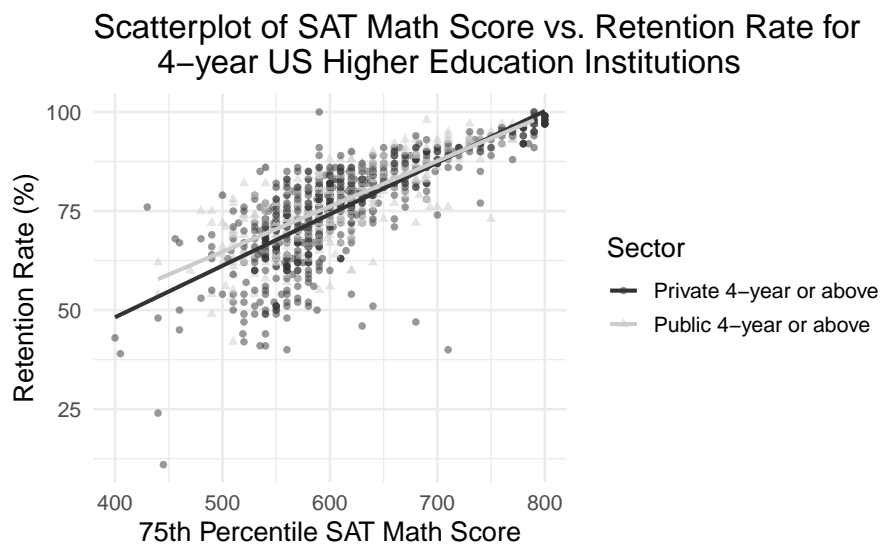
8. Calculate the value for r^2 using the values for SST and SSE provided below each of the previous graphs.

9. Write a sentence interpreting the coefficient of determination in context of the problem.

Multivariable plots

When adding another categorical predictor, we can add that variable as shape or color to the plot. In the following code we have added the variable **Sector**, whether the 4-year institution is public or private.

```
IPEDS %>% # Data sent pipes into...
  ggplot(aes(x = SATMath_75, y = Retention, shape = Sector, color=Sector))+ # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "75th Percentile SAT Math Score", # Label x-axis
       y = "Retention Rate (%)", # Label y-axis
       title = "Scatterplot of SAT Math Score vs. Retention Rate for
               4-year US Higher Education Institutions") +
  # Be sure to title your plots with the type of plot, observational units, variable(s)
  geom_smooth(method = "lm", se = FALSE) + # Add regression line
  scale_color_grey()
```



10. Does the relationship between 75th percentile SAT math score and retention rate of 4-year institutions change depending on the level of sector?

2.4.4 Take-home messages

1. The sign of correlation and the sign of the slope will always be the same. The closer the value of correlation is to -1 or $+1$, the stronger the linear relationship between the explanatory and the response variable.
2. The coefficient of determination multiplied by 100 ($r^2 \times 100$) measures the percent of variation in the response variable that is explained by the relationship with the explanatory variable. The closer the value of the coefficient of determination is to 100%, the stronger the relationship.
3. We can use the line of regression to predict values of the response variable for values of the explanatory variable. Do not use values of the explanatory variable that are outside of the range of values in the data set to predict values of the response variable (reflect on why this is true.). This is called **extrapolation**.

2.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

2.5 Activity 28: Prediction of Crocodilian Body Size

2.5.1 Learning outcomes

- Given a research question involving two quantitative variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a simulation-based hypothesis test for slope or correlation.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a slope or correlation.
- Use bootstrapping to find a confidence interval for the slope or correlation.
- Interpret a confidence interval for a slope or correlation.

2.5.2 Terminology review

In today's activity, we will use simulation-based methods for hypothesis tests and confidence intervals for a linear regression slope or correlation. Some terms covered in this activity are:

- Correlation
- Slope
- Regression line

To review these concepts, see Chapter 21 in the textbook.

2.5.3 Crocodilian Body Size

Much research surrounds using measurements of animals to estimate body-size of extinct animals. Many challenges exist in making accurate estimates for extinct crocodilians. The term crocodilians refers to all members of the family Crocodylidae ("true" crocodiles), family Alligatoridae (alligators and caimans) and family Gavialidae (gharial, Tomistoma). The researchers in this study (O'Brien 2019) state, "Among extinct crocodilians and their precursors (e.g., suchians), several methods have been developed to predict body size from suites of hard-tissue proxies. Nevertheless, many have limited applications due to the disparity of some major suchian groups and biases in the fossil record. Here, we test the utility of head width (HW) as a broadly applicable body-size estimator in living and fossil suchians." Data were collected on 76 male and female individuals of different species. Is there evidence that head width (measured in cm) is a good predictor of total body length (measured in cm) for crocodilians?

- Download the R script file from D2L and upload to the RStudio server
- Open the file and run lines 1 - 8 to load the dataset

```
# Read in data set
croc <- read.csv("https://math.montana.edu/courses/s216/data/Crocodylian_headwidth.csv")
croc <- croc %>%
  na.omit()
```

To create a scatterplot to examine the relationship between head width and total body length we will use `HW_cm` as the explanatory variable and `TL_cm` as the response variable.

- Enter the name of the explanatory and response variable in line 14
- Highlight and run lines 13 - 20

```

croc %>% # Pipe data set into...
ggplot(aes(x = explanatory, y = response))+ # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "head width (cm)", # Label x-axis
       y = "total length (cm)", # Label y-axis
       title = "Scatterplot of Crocodilian Head Width vs. Total Length") +
  # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line

```

1. Describe the features of the plot, addressing all four characteristics of a scatterplot.

If you indicated there are potential outliers, which points are they?

Hypotheses

When analyzing two quantitative variables we can either test regression slope or correlation. In both cases, we are testing that there is a linear relationship between variables.

2. Write the null hypothesis in words.
3. Write the null hypothesis to test slope in notation.
4. Write the null hypothesis to test correlation in notation.
5. Write the alternative hypothesis in words.

Summarize and visualize the data

To create the linear model output and find the value of correlation for the linear relationship...

- Enter the the name of the explanatory and response in line 25
- Highlight and run lines 25 - 27

```
#Linear model
lm.croc <- lm(response~explanatory, data=croc) #lm(response~explanatory)
round(summary(lm.croc)$coefficients, 5)
#Correlation
cor(croc$HW_cm, croc$TL_cm)
```

6. Using the output from the evaluated R code, write the equation of the regression line in the context of the problem using appropriate statistical notation.

7. Interpret the estimated slope in context of the problem.

Use statistical inferential methods to draw inferences from the data

In this activity, we will focus on using simulation-based methods for inference in regression.

Simulation-based hypothesis test

Let's start by thinking about how one simulation would be created on the null distribution using cards. First, we would write the values for the response variable, total length, on each card. Next, we would shuffle these y values while keeping the x values (explanatory variable) in the same order. Then, find the line of regression for the shuffled (x, y) pairs and calculate either the slope or correlation of the shuffled sample.

We will use the `regression_test()` function in R (in the `catstats` package) to simulate the null distribution of shuffled slopes (or shuffled correlations) and compute a p-value. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `croc`), the summary measure for the test (either slope or correlation), number of repetitions, the sample statistic (value of slope or correlation), and the direction of the alternative hypothesis.

The response variable name is `TL_cm` and the explanatory variable name is `HW_cm` for these data.

8. What inputs should be entered for each of the following to create the simulation to test regression slope?
 - Direction ("`greater`", "`less`", or "`two-sided`"):
 - Summary measure (choose "`slope`" or "`correlation`"):
 - As extreme as (enter the value for the sample slope):

- Number of repetitions:

Using the R script file for this activity...

- Enter your answers for question 8 in place of the xx's to produce the null distribution with 1000 simulations.
- Highlight and run lines 32–37.

```
regression_test(TL_cm~HW_cm, # response ~ explanatory
               data = croc, # Name of data set
               direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
               summary_measure = "xx", # "slope" or "correlation"
               as_extreme_as = xx, # Observed slope or correlation
               number_repetitions = 10000) # Number of simulated samples for null distribution
```

9. Report the p-value from the R output.
10. Suppose we wanted to complete the simulation test using correlation as the summary measure, instead of slope. Which two inputs in #8 would need to be changed to test for correlation? What inputs should you use instead?
11. Change the inputs in lines 32–37 to test for correlation instead of slope. Highlight and run those lines, then report the new p-value of the test.
12. The p-values from the test of slope (#9) and the test of correlation (#11) should be similar. Explain why the two p-values should match. *Hint: think about the relationship between slope and correlation!*

Simulation-based confidence interval

We will use the `regression_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample slopes (or sample correlations) and calculate a confidence interval.

- Fill in the missing values in the provided R script file to find a 95% confidence interval for slope.
- Highlight and run lines 42–46.

```
regression_bootstrap_CI(response~explanatory, # response ~ explanatory
                       data = croc, # Name of data set
                       confidence_level = xx, # Confidence level as decimal
                       summary_measure = "xx", # Slope or correlation
                       number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```

13. Report the bootstrap 95% confidence interval in interval notation.

14. Interpret the interval in question 14 in context of the problem. *Hint: use the interpretation of slope in your confidence interval interpretation.*

Communicate the results and answer the research question

15. Based on the p-value and confidence interval, write a conclusion in context of the problem.

2.5.4 Take-home messages

1. The p-value for a test for correlation should be approximately the same as the p-value for the test of slope. In the simulation test, we just change the statistic type from slope to correlation and use the appropriate sample statistic value.
2. To interpret a confidence interval for the slope, think about how to interpret the sample slope and use that information in the confidence interval interpretation for slope.
3. To create one simulated sample on the null distribution when testing for a relationship between two quantitative variables, hold the x values constant and shuffle the y values to new x values. Find the regression line for the shuffled data and plot the slope or the correlation for the shuffled data.
4. To create one simulated sample on the bootstrap distribution when assessing two quantitative variables, label n cards with the original (response, explanatory) values. Randomly draw with replacement n times. Find the regression line for the resampled data and plot the resampled slope or correlation.

2.5.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

2.6 Activity 29: Golf Driving Distance

2.6.1 Learning outcomes

- Given a research question involving two quantitative variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Assess the conditions to use the normal distribution model for a slope.
- Find the T test statistic (T-score) for a slope based off of `lm()` output in R.
- Find, interpret, and evaluate the p-value for a theory-based hypothesis test for a slope.
- Create and interpret a theory-based confidence interval for a slope.
- Use a confidence interval to determine the conclusion of a hypothesis test.

2.6.2 Terminology review

In this week's in-class activity, we will use theory-based methods for hypothesis tests and confidence intervals for a linear regression slope. Some terms covered in this activity are:

- Slope
- Regression line

To review these concepts, see Chapter 21 in the textbook.

2.6.3 Golf driving distance

In golf the goal is to complete a hole with as few strokes as possible. A long driving distance to start a hole can help minimize the strokes necessary to complete the hole, as long as that drive stays on the fairway. Data were collected on 354 PGA and LPGA players in 2008 ("Average Driving Distance and Fairway Accuracy" 2008). For each player, the average driving distance (yards), fairway accuracy (percentage), and sex was measured. Use these data to assess, "Does a professional golfer give up accuracy when they hit the ball farther?"

- Download the R script file from D2L and open in the RStudio server

```
# Read in data set
golf <- read.csv("https://math.montana.edu/courses/s216/data/golf.csv")
```

Plot review.

To create a scatterplot showing the relationship between the driving distance and percent accuracy for professional golfers:

- Enter the name of the explanatory and response in line 10
- Highlight and run lines 1 - 16

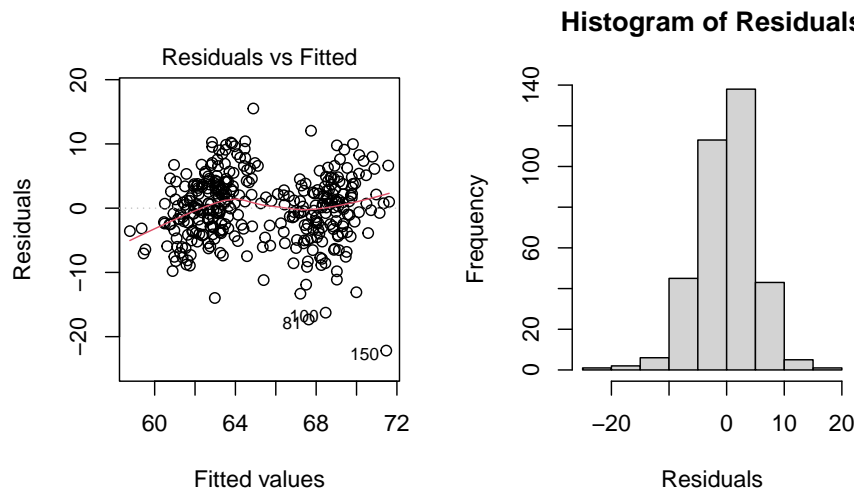
```
golf %>% # Pipe data set into...
ggplot(aes(x = explanatory, y = response))+ # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "Driving Distance (yards)", # Label x-axis
       y = "Percent Accuracy", # Label y-axis
       title = "Scatterplot of Driving Distance by Percent Accuracy
for Professional Golfers") +
  # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line
```

Conditions for the least squares line

When performing inference on a least squares line, the follow conditions are generally required:

- *Independent observations* (for both simulation-based and theory-based methods): individual data points must be independent.
 - Check this assumption by investigating the sampling method and determining if the observational units are related in any way.
- *Linearity* (for both simulation-based and theory-based methods): the data should follow a linear trend.
 - Check this assumption by examining the scatterplot of the two variables, and a scatterplot of the residuals (on the y -axis) versus the fitted values (on the x -axis). The pattern in the residual plot should display a horizontal line.
- *Constant variability* (for theory-based methods only): the variability of points around the least squares line remains roughly constant
 - Check this assumption by examining a scatterplot of the residuals (on the y -axis) versus the fitted values (on the x -axis). The variability in the residuals around zero should be approximately the same for all fitted values.
- *Nearly normal residuals* (for theory-based methods only): residuals must be nearly normal.
 - Check this assumption by examining a histogram of the residuals, which should appear approximately normal.

The scatterplot generated earlier and the residual plots shown below will be used to assess these conditions for approximating the data with the t -distribution.



1. Are the conditions met to use the t -distribution to approximate the sampling distribution of the standardized statistic? Justify your answer.

Ask a research question

2. Write out the null hypothesis in words to test the slope.
3. Using the research question, write the alternative hypothesis in notation to test the slope.

Summarize and visualize the data

The linear model output for this study is shown below.

```
lm.golf <- lm(Percent_Accuracy~Driving_Distance, data=golf) # lm(response~explanatory)
round(summary(lm.golf)$coefficients, 3)
```

```
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      103.586      3.329   31.119      0
#> Driving_Distance  -0.142      0.012  -11.553      0
```

4. Report the summary statistic (sample slope) for the linear relationship between driving distance and percent accuracy of golfers. Use proper notation.

Use statistical inferential methods to draw inferences from the data

Hypothesis test To find the value of the standardized statistic to test the slope we will use,

$$T = \frac{\text{slope estimate} - \text{nullvalue}}{SE} = \frac{b_1 - 0}{SE(b_1)}.$$

We will use the linear model R output above to get the estimate for slope and the standard error of the slope.

5. Calculate the standardized statistic for slope. Identify where this calculated value is in the linear model R output.
6. The p-value in the linear model R output is the two-sided p-value for the test of significance for slope. Report the p-value to answer the research question.
7. Based on the p-value, how much evidence is there against the null hypothesis?

Confidence interval Recall that a confidence interval is calculated by adding and subtracting the margin of error to the point estimate.

$$\text{point estimate} \pm t^* \times SE(\text{estimate}).$$

When the point estimate is a regression slope, this formula becomes

$$b_1 \pm t^* \times SE(b_1).$$

The t^* multiplier comes from a t -distribution with $n - 2$ degrees of freedom. The sample size for this study is 354 so we will use the degrees of freedom 352 ($n - 2$).

- Enter the percentile needed to find the multiplier for a 95% confidence interval for xx
- Enter the degrees of freedom for yy
- Highlight and run line 33

```
qt(xx, yy, lower.tail = TRUE) # 95% t* multiplier
```

8. Calculate the 95% confidence interval for the true slope.

9. Interpret the 95% confidence interval in context of the problem.

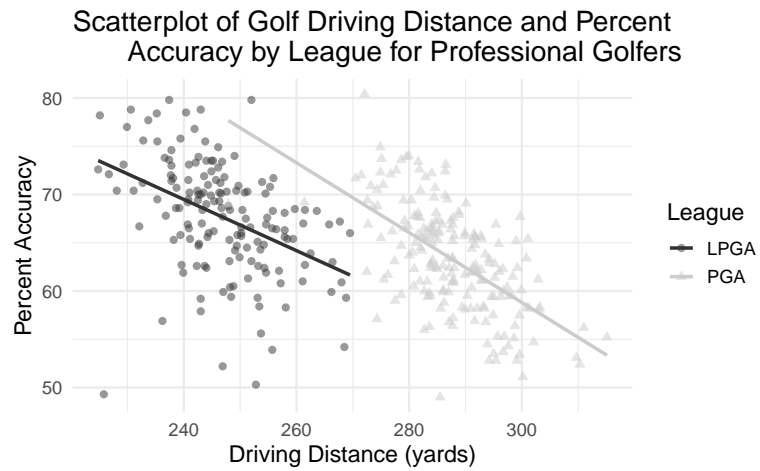
Communicate the results and answer the research question

10. Write a conclusion to answer the research question in context of the problem.

Multivariable plots

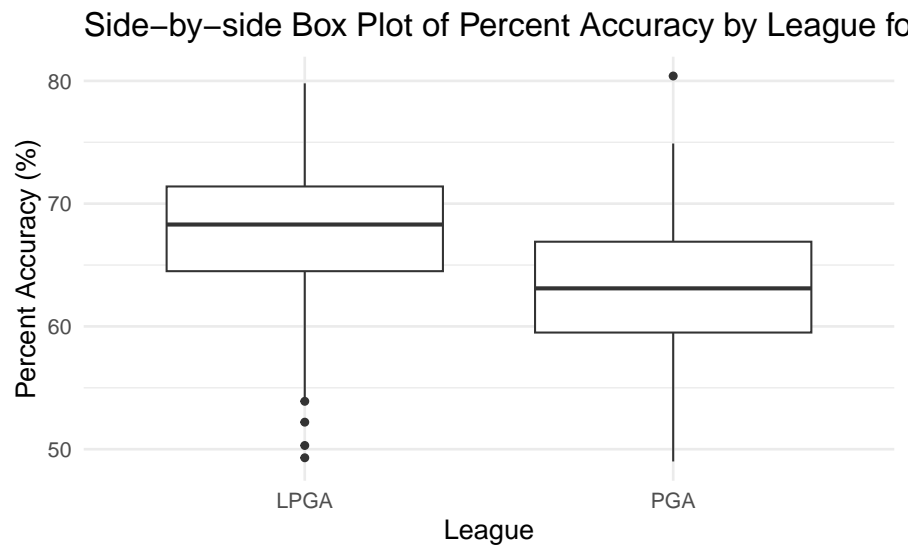
Another variable that may affect the percent accuracy is the which league the golfer is part of. We will look at how this variable may change the relationship between driving distance and percent accuracy.

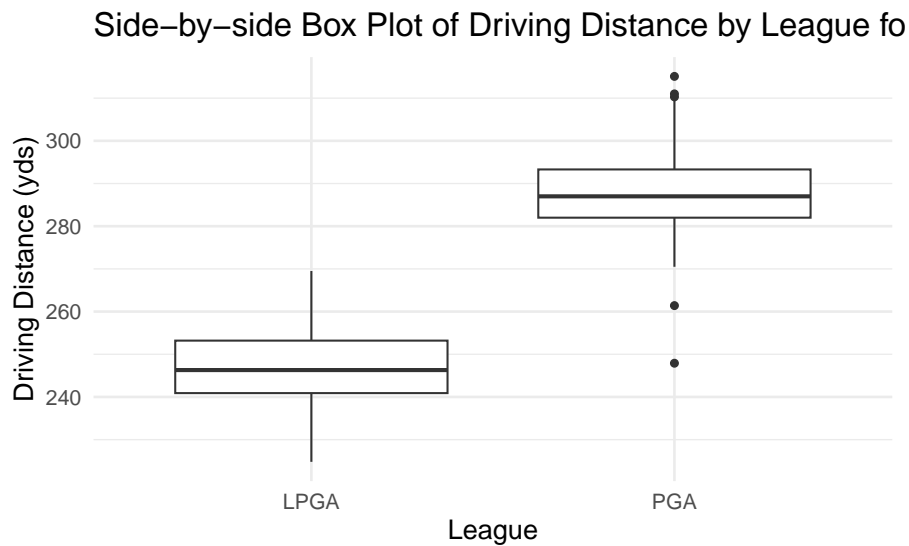
```
golf %>%  
  ggplot(aes(x = Driving_Distance, y = Percent_Accuracy, color=League))+ # Specify variables  
  geom_point(aes(shape = League), size = 2, alpha=0.5) + # Add scatterplot of points  
  labs(x = "Driving Distance (yards)", # Label x-axis  
       y = "Percent Accuracy", # Label y-axis  
       color = "League", shape = "League",  
       title = "Scatterplot of Golf Driving Distance and Percent  
Accuracy by League for Professional Golfers") + # Be sure to title your plots  
  geom_smooth(method = "lm", se = FALSE) + # Add regression line  
  scale_color_grey()
```



11. Does the association between driving distance and percent accuracy change depending on which league the golfer is a part of? Explain your answer.

12. Explain the association between league and each of the other two variables. Use the following plots in addition to the scatterplot from Q9 to explain your answer.





2.6.4 Take-home messages

1. To check the validity conditions for using theory-based methods we must use the residual diagnostic plots to check for normality of residuals and constant variability, and the scatterplot to check for linearity.
2. To interpret a confidence interval for the slope, think about how to interpret the sample slope and use that information in the confidence interval interpretation for slope.
3. Use the explanatory variable row in the linear model R output to obtain the slope estimate (**estimate** column) and standard error of the slope (**Std. Error** column) to calculate the standardized slope, or T-score. The calculated T-score should match the **t value** column in the explanatory variable row. The standardized slope tells the number of standard errors the observed slope is above or below 0.
4. The explanatory variable row in the linear model R output provides a **two-sided** p-value under the **Pr(>|t|)** column.
5. The standardized slope is compared to a t -distribution with $n - 2$ degrees of freedom in order to obtain a p-value. The t -distribution with $n - 2$ degrees of freedom is also used to find the appropriate multiplier for a given confidence level.

2.6.5 Additional notes

Use this space to summarize your thoughts and take additional notes on this week's activity and material covered.

2.7 Module 13 Lab: Big Mac Index

2.7.1 Learning outcomes

- Given a research question involving two quantitative variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Assess the conditions to determine in theory or simulation-based methods should be used.
- Find, interpret, and evaluate the p-value for a hypothesis test for a slope or correlation.
- Create and interpret a confidence interval for a slope or correlation.

2.7.2 Big Mac Index

Can the relative cost of a Big Mac across different countries be used to predict the Gross Domestic Product (GDP) per person for that country? The log GDP per person and the adjusted dollar equivalent to purchase a Big Mac was found on a random sample of 55 countries in January of 2022. The cost of a Big Mac in each country was adjusted to US dollars based on current exchange rates. Is there evidence of a positive relationship between Big Mac cost (`dollar_price`) and the log GDP per person (`log_GDP`)?

- Upload and open the R script file for Week 13 lab.
- Upload the csv file, `big_mac_adjusted_index_S22.csv`.
- Enter the name of the data set for `datasetname` in the R script file in line 9.
- Highlight and run lines 1–9 to load the data.

```
# Read in data set
mac <- read.csv("datasetname")
```

Summarize and visualize the data

- To find the correlation between the variables, `log_GDP` and `dollar_price` highlight and run lines 13–16 in the R script file.

```
mac %>%
  select(c("log_GDP", "dollar_price")) %>%
  cor(use="pairwise.complete.obs") %>%
  round(3)
```

1. Report the value of correlation between the variables.
2. Calculate the value of the coefficient of determination between `log_GDP` and `dollar_price`.
3. Interpret the value of the coefficient of determination in context of the problem.

In the next part of the activity we will assess the linear model between Big Mac cost and log GDP.

- Enter the variable `log_GDP` for **response** and the variable `dollar_price` for **explanatory** in line 22.
- Highlight and run lines 22–23 to get the linear model output.

```
# Fit linear model: y ~ x
bigmacLM <- lm(response~explanatory, data=mac)
round(summary(bigmacLM)$coefficients,3) # Display coefficient summary
```

4. Give the value of the slope of the regression line. Interpret this value in context of the problem.

Conditions for the least squares line

5. Is there independence between the responses for the observational units? Justify your answer.

- Highlight and run lines 28–33 to create the scatterplot to check for linearity.

```
#Scatterplot
mac %>% # Pipe data set into...
  ggplot(aes(x = dollar_price, y = log_GDP))+ # Specify variables
  geom_point(alpha=0.5) + # Add scatterplot of points
  labs(x = "Big Mac Cost", # Label x-axis
       y = "log GDP", # Label y-axis
       title = "Scatterplot of Big Mac Cost vs. log GDP per person  
for Countries in 2022") + # Be sure to tile your plots
  geom_smooth(method = "lm", se = FALSE) # Add regression line
```

6. Is the linearity condition met to use regression methods to analyze the data? Justify your answer.

- Highlight and run lines 38–42 to produce the diagnostic plots needed to assess conditions to use theory-based methods.

```
#Diagnostic plots
bigmacLM <- lm(log_GDP~dollar_price, data = mac) # Fit linear regression model
par(mfrow=c(1,2)) # Set graphics parameters to plot 2 plots in 1 row
plot(bigmacLM, which=1) # Residual vs fitted values
hist(bigmacLM$resid, xlab="Residuals", ylab="Frequency",
     main = "Histogram of Residuals") # Histogram of residuals
```

7. Are the conditions met to use the t -distribution to approximate the sampling distribution of the standardized statistic? Justify your answer.

Ask a research question

8. Write out the null and alternative hypotheses in notation to test *correlation* between Big Mac cost and country GDP.

H_0 :

H_A :

Use statistical inferential methods to draw inferences from the data

Hypothesis test

Use the `regression_test()` function in R (in the `catstats` package) to simulate the null distribution of sample **correlations** and compute a p-value. We will need to enter the response variable name and the explanatory variable name for the formula, the data set name (identified above as `mac`), the summary measure used for the test, number of repetitions, the sample statistic (value of correlation), and the direction of the alternative hypothesis.

The response variable name is `log_GDP` and the explanatory variable name is `dollar_price`.

9. What inputs should be entered for each of the following to create the simulation to test correlation?

- Direction ("**greater**", "**less**", or "**two-sided**"):
- Summary measure (choose "**slope**" or "**correlation**"):
- As extreme as (enter the value for the sample correlation):
- Number of repetitions:

Using the R script file for this activity, enter your answers for question 9 in place of the `xx`'s to produce the null distribution with 10000 simulations.

- Highlight and run lines 47–53.
- Upload a copy of your plot showing the p-value to Gradescope for your group.

```
regression_test(log_GDP~dollar_price, # response ~ explanatory
  data = mac, # Name of data set
  direction = "xx", # Sign in alternative ("greater", "less", "two-sided")
  summary_measure = "xx", # "slope" or "correlation"
  as_extreme_as = xx, # Observed slope or correlation
  number_repetitions = 10000) # Number of simulated samples for null distribution
```

10. Report the p-value from the R output.

Simulation-based confidence interval

We will use the `regression_bootstrap_CI()` function in R (in the `catstats` package) to simulate the bootstrap distribution of sample **correlations** and calculate a confidence interval.

- Fill in the `xx`'s in the the provided R script file to find a 90% confidence interval.
- Highlight and run lines 58–62.

```
regression_bootstrap_CI(log_GDP~dollar_price, # response ~ explanatory
  data = mac, # Name of data set
  confidence_level = xx, # Confidence level as decimal
  summary_measure = "xx", # Slope or correlation
  number_repetitions = 10000) # Number of simulated samples for bootstrap distribution
```

11. Report the bootstrap 90% confidence interval in interval notation.

Communicate the results and answer the research question

12. Using a significance level of 0.1, what decision would you make?
13. What type of error is possible?
14. Interpret this error in context of the problem.
15. Write a paragraph summarizing the results of the study as if you are reporting these results in your local newspaper. **Upload a copy of your paragraph to Gradescope for your group.** Be sure to describe:
 - Summary statistic and interpretation
 - Summary measure (in context)
 - Value of the statistic
 - Order of subtraction when comparing two groups
 - P-value and interpretation
 - Statement about probability or proportion of samples
 - Statistic (summary measure and value)
 - Direction of the alternative
 - Null hypothesis (in context)
 - Confidence interval and interpretation
 - How confident you are (e.g., 90%, 95%, 98%, 99%)
 - Parameter of interest

- Calculated interval
 - Order of subtraction when comparing two groups
- Conclusion (written to answer the research question)
 - Amount of evidence
 - Parameter of interest
 - Direction of the alternative hypothesis
- Scope of inference
 - To what group of observational units do the results apply (target population or observational units similar to the sample)?
 - What type of inference is appropriate (causal or non-causal)?

- “Average Driving Distance and Fairway Accuracy.” 2008. <https://www.pga.com/> and <https://www.lpga.com/>.
- Banton, et al, S. 2022. “Jog with Your Dog: Dog Owner Exercise Routines Predict Dog Exercise Routines and Perception of Ideal Body Weight.” *PLoS ONE* 17(8).
- Bhavsar, et al, A. 2022. “Increased Risk of Herpes Zoster in Adults ≥ 50 Years Old Diagnosed with COVID-19 in the United States.” *Open Forum Infectious Diseases* 9(5).
- Bulmer, M. n.d. “Islands in Schools Project.” <https://sites.google.com/site/islandsinschoolsprojectwebsite/home> e.
- “Bureau of Transportation Statistics.” 2019. <https://www.bts.gov/>.
- “Child Health and Development Studies.” n.d. <https://www.chdstudies.org/>.
- Darley, J. M., and C. D. Batson. 1973. “From Jerusalem to Jericho”: A Study of Situational and Dispositional Variables in Helping Behavior.” *Journal of Personality and Social Psychology* 27: 100–108.
- Davis, Smith, A. K. 2020. “A Poor Substitute for the Real Thing: Captive-Reared Monarch Butterflies Are Weaker, Paler and Have Less Elongated Wings Than Wild Migrants.” *Biology Letters* 16.
- Du Toit, et al, G. 2015. “Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy.” *New England Journal of Medicine* 372.
- Edmunds, et al, D. 2016. “Chronic Wasting Disease Drives Population Decline of White-Tailed Deer.” *PLoS ONE* 11(8).
- Education Statistics, National Center for. 2018. “IPEDS.” <https://nces.ed.gov/ipeds/>.
- “Great Britain Married Couples: Great Britain Office of Population Census and Surveys.” n.d. <https://discovery.nationalarchives.gov.uk/details/r/C13351>.
- Group, TODAY Study. 2012. “A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes.” *New England Journal of Medicine* 366: 2247–56.
- Hamblin, J. K., K. Wynn, and P. Bloom. 2007. “Social Evaluation by Preverbal Infants.” *Nature* 450 (6288): 557–59.
- Hirschfelder, A., and P. F. Molin. 2018. “I Is for Ignoble: Stereotyping Native Americans.” Retrieved from <https://www.ferris.edu/HTMLS/news/jimcrow/native/homepage.htm>.
- Hutchison, R. L., and M. A. Hirthler. 2013. “Upper Extremity Injuries in Homer’s Iliad.” *Journal of Hand Surgery (American Volume)* 38: 1790–93.
- “IMDb Movies Extensive Dataset.” 2016. <https://kaggle.com/stefanoleone992/imdb-extensive-dataset>.
- Kalra, et al., D. 2022. “Trustworthiness of Indian Youtubers.” Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/4426566>.
- Keating, D., N. Ahmed, F. Nirappil, Stanley-Becker I., and L. Bernstein. 2021. “Coronavirus Infections Dropping Where People Are Vaccinated, Rising Where They Are Not, Post Analysis Finds.” *Washington Post*. <https://www.washingtonpost.com/health/2021/06/14/covid-cases-vaccination-rates/>.
- Laeng, Mathisen, B. 2007. “Why Do Blue-Eyed Men Prefer Women with the Same Eye Color?” *Behavioral Ecology and Sociobiology* 61(3).
- Levin, D. T. 2000. “Race as a Visual Feature: Using Visual Search and Perceptual Discrimination Tasks to Understand Face Categories and the Cross-Race Recognition Deficit.” *Journal of Experimental Psychology* 129(4).
- LUETKEMEIER, et al., M. 2017. “Skin Tattoos Alter Sweat Rate and Na⁺ Concentration.” *Medicine and Science in Sports and Exercise* 49(7).
- Madden, et al, J. 2020. “Ready Student One: Exploring the Predictors of Student Learning in Virtual Reality.” *PLoS ONE* 15(3).
- Miller, G. A. 1956. “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information.” *Psychological Review* 63(2).
- Moquin, W., and C. Van Doren. 1973. “Great Documents in American Indian History.” Praeger.
- “More Americans Are Joining the ‘Cashless’ Economy.” 2022. <https://www.pewresearch.org/short-reads/2022/10/05/more-americans-are-joining-the-cashless-economy/>.
- National Weather Service Corporate Image Web Team. n.d. “National Weather Service – NWS Billings.” <https://w2.weather.gov/climate/xmacis.php?wfo=byz>.
- O’Brien, Lynch, H. D. 2019. “Crocodylian Head Width Allometry and Phylogenetic Prediction of Body Size in Extinct Crocodyliforms.” *Integrative Organismal Biology* 1.
- “Ocean Temperature and Salinity Study.” n.d. <https://calcofi.org/>.

- “Older People Who Get Covid Are at Increased Risk of Getting Shingles.” 2022. <https://www.washingtonpost.com/health/2022/04/19/shingles-and-covid-over-50/>.
- “Physician’s Health Study.” n.d. <https://phs.bwh.harvard.edu/>.
- Porath, Erez, C. 2017. “Does Rudeness Really Matter? The Effects of Rudeness on Task Performance and Helpfulness.” *Academy of Management Journal* 50.
- Quinn, G. E., C. H. Shin, M. G. Maguire, and R. A. Stone. 1999. “Myopia and Ambient Lighting at Night.” *Nature* 399 (6732): 113–14. <https://doi.org/10.1038/20094>.
- Ramachandran, V. 2007. “3 Clues to Understanding Your Brain.” https://www.ted.com/talks/vs_ramachandran_3_clues_to_understanding_your_brain.
- “Rates of Laboratory-Confirmed COVID-19 Hospitalizations by Vaccination Status.” 2021. CDC. <https://covid.cdc.gov/covid-data-tracker/#covidnet-hospitalizations-vaccination>.
- Richardson, T., and R. T. Gilman. 2019. “Left-Handedness Is Associated with Greater Fighting Success in Humans.” *Scientific Reports* 9 (1): 15402. <https://doi.org/10.1038/s41598-019-51975-3>.
- Stephens, R., and O. Robertson. 2020. “Swearing as a Response to Pain: Assessing Hypoalgesic Effects of Novel “Swear” Words.” *Frontiers in Psychology* 11: 643–62.
- Stewart, E. H., B. Davis, B. L. Clemans-Taylor, B. Littenberg, C. A. Estrada, and R. M. Centor. 2014. “Rapid Antigen Group a Streptococcus Test to Diagnose Pharyngitis: A Systematic Review and Meta-Analysis” 9 (11). <https://doi.org/10.1371/journal.pone.0111727>.
- Stroop, J. R. 1935. “Studies of Interference in Serial Verbal Reactions.” *Journal of Experimental Psychology* 18: 643–62.
- Subach, et al, A. 2022. “Foraging Behaviour, Habitat Use and Population Size of the Desert Horned Viper in the Negev Desert.” *Soc. Open Sci* 9.
- Sulheim, S., A. Ekeland, I. Holme, and R. Bahr. 2017. “Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders: Changes After an Interval of One Decade” 51 (1): 44–50. <https://doi.org/10.1136/bjsports-2015-095798>.
- “Titanic.” n.d. <http://www.encyclopedia-titanica.org>.
- “US COVID-19 Vaccine Tracker: See Your State’s Progress.” 2021. Mayo Clinic. <https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker>.
- US Environmental Protection Agency. n.d. “Air Data – Daily Air Quality Tracker.” <https://www.epa.gov/outdoor-air-quality-data/air-data-daily-air-quality-tracker>.
- Wahlstrom, et al, K. 2014. “Examining the Impact of Later School Start Times on the Health and Academic Performance of High School Students: A Multi-Site Study.” *Center for Applied Research and Educational Improvement*.
- Watson, et al., N. 2015. “Recommended Amount of Sleep for a Healthy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society.” *Sleep* 38(6).
- Weiss, R. D. 1988. “Relapse to Cocaine Abuse After Initiating Desipramine Treatment.” *JAMA* 260(17).
- “Welcome to the Navajo Nation Government: Official Site of the Navajo Nation.” 2011. Retrieved from <https://www.navajo-nsn.gov/>.
- Wilson, Woodruff, J. P. 2016. “Vertebral Adaptations to Large Body Size in Theropod Dinosaurs.” *PLoS ONE* 11(7).