STAT 216 Coursepack

Fall 2025
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

Preface

Placeholder

MODULE 1

Basics of Data and Sampling Methods

Placeholder

1.1 Vocabulary Review and Key Topics

1.1.1 Key topics

1.1.2 Vocabulary

1.2 Video Notes: Intro to data and Sampling Methods

1.2.1 Course Videos

Data basics: Video 1.2.1and1.2.2

Types of variables

Exploratory data analysis (EDA)

Sampling Methods: Video 2.1

Good vs. bad sampling

Types of Sampling Bias

Optional Notes: Video Example

1.2.2 Concept Check

1.3 Activity 1: Intro to Data Analysis and Sampling Bias

1.3.1 Learning outcomes

1.3.2 Terminology review

Notes on Observational Units and Variables

Further analysis of class data set

Notes on Sampling Methods and Types of bias

Types of bias

1.3.3 Take-home messages

1.3.4 Additional notes

1.4 Activity 2: American Indian Address

1.4.1 Learning outcomes

1.4.2 Terminology review

1.4.3 Class Preparation

1.4.4 American Indian Address

By eye selection

Notes on sampling

1.4.5 Class Activity

Random selection

Effect of sample size

1.4.6 Take-home messages

1.4.7 Additional notes

MODULE 2

Probability

Placeholder

2.1 Vocabulary Review and Key Topics

2.1.1 Key topics

2.1.2 Vocabulary

2.2 Video Notes: Probability

2.2.1 Course Videos

Probability

Creating a hypothetical two-way table

Diagnostic tests

2.2.2 Concept Check

2.3 Activity 3: Probability Studies

2.3.1 Learning outcomes

2.3.2 Terminology review

Notes on probability

Probability notation

Probability questions

Calculating probabilities from a two-way table

2.3.3 Take home messages

2.3.4 Additional notes

MODULE 3

Exploring Categorical Data: Exploratory Data Analysis and Inference using Simulation-based Methods

Placeholder

3.1 Vocabulary Review and Key Topics

3.1.1 Key topics

Steps of the statistical investigation process

3.1.2 Vocabulary

Plotting one categorical variable

Inference

Simulation-based inference for a single proportion

3.2 Video Notes: Exploratory Data Analysis of Categorical Variables

3.2.1 Course Videos

Summarizing categorical data - Video 4.1_OneProp

Optional Notes: Video Example

Displaying categorical variables - Video 4.2_OneProp

Hypothesis Testing - Video Chapter9

Hypothesis Testing/Justice System

Hypotheses

Null hypothesis

Alternative hypothesis

Simulation vs. Theory-based Methods

Simulation-based method

Theory-based method

P-value

Hypothesis testing

Confidence interval - Video Chapter10

Sampling distribution

Simulation-based methods

Optional Notes: Video Example (Video 14.1)

Simulation-based method

Optional Notes: Video Example (Video 14.2)

3.2.2 Concept Check

3.3 Activity 4: Helper-Hinderer Part 1 — Simulation-based Hypothesis Test

3.3.1 Learning outcomes

3.3.2 Terminology review

3.3.3 Steps of the statistical investigation process

Notes on one categorical variable

MODULE 4

Inference for a Single Categorical Variable: Theory-based Methods

Placeholder

4.1 Vocabulary Review and Key Topics

4.1.1 Key topics

4.1.2 Vocabulary

4.2 Video Notes: Inference for One Categorical Variable using Theory-based Methods

4.2.1 Course Videos

Theory-based methods

Central limit theorem - Video Chapter11

68-95-99.7 Rule

Theoretical Testing for a Single Proportion - Video 14.3TheoryTests

Optional Notes: Video Example (Video 14.3TheoryTests)

Theoretical Confidence Intervals for a Single Proportion - Video 14.3TheoryIntervals

Theory-based method for a single categorical variable

Optional Notes: Video Example (Video 14.3TheoryIntervals)

Interpreting confidence level - Video 14.3TheoryIntervals

4.2.2 Concept Check

4.3 Activity 7: Handedness of Male Boxers

4.3.1 Learning outcomes

4.3.2 Terminology review

4.3.3 Handedness of male boxers

R Instructions

Hypotheses and summary statistics

Theory-based methods

Additional notes on Theory-based methods

Impacts on the p-value

4.3.4 Take-home messages

4.3.5 Additional notes

4.4 Activity 8: Confidence intervals and what confidence means

4.4.1 Learning outcomes

4.4.2 Terminology review

4.4.3 Handedness of male boxers continued

What does *confidence* mean?

Notes on theory-based confidence intervals

Simulation methods

Effect of sample size on the width of the confidence interval

4.4.4 Take-home messages

4.4.5 Additional notes

4.5 Module 3 and 4 Lab: Mixed Breed Dogs in the U.S.

4.5.1 Learning outcomes

4.5.2 Mixed Breed Dogs in the U.S.

R Instructions

Use statistical analysis methods to draw inferences from the data

Null Distribution

Bootstrap distribution

Summarize the results of the study

MODULE 5

Unit 1 Review

The following module contains both a list of key topics covered in Unit 1 as well as Module Review Worksheets that will be covered in Weekly Review Sessions.

5.0.1 Key Topics

Review the key topics for Unit 1 prior to the first exams. All of these topics will be covered in Modules 1–4.

5.0.2 Module Review

The following worksheets review each of the modules. These worksheets will be completed during Melinda's Study Sessions each week. Solutions will be posted on Canvas in the Unit 1 Review folder after the study sessions.

5.1 Key Topics Exam 1

Descriptive statistics and study design

Hypothesis testing

5.1.1 Confidence intervals

5.1.2 Probability

5.2 Module 1 Review - Sampling Methods

5.3 Module 2 Review - Probability

5.4 Module 3 Review - Simulation Methods for a Single Proportion

5.5 Module 4 Review - Theory-based Methods for a Single Proportion

5.6 Group Exam 1 Review

MODULE 6

Exploring Quantitative Data: Exploratory Data Analysis and Inference for a Single Quantitative Variable - Simulation-based Methods

Placeholder

6.1 Vocabulary Review and Key Topics

6.1.1 Key topics

6.1.2 Vocabulary

Sample statistics for a single quantitative variable

Plotting one quantitative variable

Hypothesis testing for a single mean

Simulation-based hypothesis testing

Simulation-based confidence interval

6.2 Video Notes: Exploratory Data Analysis and Hypothesis Testing of Quantitative Variables

6.2.1 Course Videos

Summarizing quantitative data - Video 5.2to5.4

Types of plots

Summarizing quantitative data - Video 5.5

Four characteristics of plots for quantitative variables

Robust statistics - Video 5.7

Simulation-based Testing for a Single Mean - Video 17.2

Hypothesis testing

Simulation-based method

Optional Notes: Video Example (Video 17.2)

Simuation-based Confidence Intervals for a Single Mean - Video 17.1

Confidence interval

Simulation-based method

6.2.2 Concept Check

6.3 Activity 9: Summarizing Quantitative Variables

6.3.1 Learning outcomes

6.3.2 Terminology review

6.3.3 The Integrated Postsecondary Education Data System (IPEDS)

Identifying variables in a data set

Notes on Summarizing Quantitative Variables:

R Instructions

Displaying a single quantitative variable

Robust statistics

6.3.4 Take-home messages

6.3.5 Additional notes

MODULE 7

Exploring Quantitative Data: Inference for a Single Quantitative Variable - Theory-based Methods

Placeholder

7.1 Vocabulary Review and Key Topics

7.1.1 Key topics

Theory-based hypothesis testing

Theory-based confidence interval

Vocabulary

7.2 Video Notes: Theory-based Inference for a single quantitative variable

7.2.1 Course Videos

Theory-based Testing for a Single Mean - Video 17.3TheoryTests

t-distribution

Optional Notes: Video Example (Video 17.3TheoryTests)

Theory-based Confidence Interval for a Single Mean - Video 17.3TheoryIntervals

Decisions, Errors, and Power - Video Chapter12

7.2.2 Concept Check

7.3 Activity 11: Body Temperature

7.3.1 Learning outcomes

7.3.2 Terminology review

7.3.3 Body Temperature

Ask a research question

Summarize and visualize the data

Check theoretical conditions

Use statistical inferential methods to draw inferences from the data

Theory-based methods to create a confidence interval

7.3.4 Take-home messages

7.3.5 Additional notes

7.4 Activity 12: Errors and Power

7.4.1 Learning outcomes

7.4.2 Terminology review

Notes on types of errors and power

7.4.3 College textbook cost

7.4.4 Take-home messages

7.4.5 Additional notes

7.5 Module 6 and 7 Lab: Arsenic

7.5.1 Learning outcomes

MODULE 8

Exploratory Data Analysis and Simulation-based Inference for Two Categorical Variables

Placeholder

8.1 Vocabulary Review and Key Topics

8.1.1 Key topics

8.1.2 Vocabulary

Sample statistics for two categorical variables

Plotting two categorical variables

Hypotheses

Simulation-based hypothesis testing for a difference in proportions

Simulation-based confidence interval

Study design

Scope of inference

8.2 Video Notes: Inference for Two Categorical Variables using Simulation-based Methods

8.2.1 Course Videos

Relationships between variables - Video 1.2.3to1.2.5

Relationships between variables

Observational studies, experiments, and scope of inference: Video 2.2to2.4

Study design

Optional Notes: Video Examples (Video 2.2to2.4)

Scope of Inference

Summarizing two categorical variables - Video 4.1_TwoProp

Plots for two categorical variables - Video 4.2_TwoProp

Simpson's paradox - Video 4.4

Simulation Testing for a Difference in Proportions - Video 15.1

Hypothesis Testing

Optional Notes: Video Example (Video 15.1)

Summary statistics and plot

Simulation-based method

Confidence interval for a Difference in Proportion - Video 15.2

Simulation-based method

Optional Notes: Video Example (Video 15.2)

Relative Risk - Video RelativeRisk

Optional Notes: Video Example (Video RelativeRisk)

Relative risk in the news

Testing Relative Risk

8.2.2 Concept Check

MODULE **9**

Theory-based Hypothesis Testing and Confidence Intervals for Two Categorical Variables:

Placeholder

9.1 Vocabulary Review and Key Topics

9.1.1 Key topics

9.1.2 Vocabulary

Theory-based inference

9.2 Video Notes: Theoretical Inference for Two Categorical Variables

9.2.1 Course Videos

Theoretical Testing for a Difference in Proportion - Video 15.4TheoryTests

Optional Notes: Video Example (Video 15.3TheoryTests)

Theoretical Confidence Interval for a Difference in Proportion - Video 15.3TheoryIntervals

Theory-based method for a two categorical variables

Optional Notes: Video Example (Video 15.3TheoryIntervals)

9.2.2 Concept Check

9.3 Activity 16: Winter Sports Helmet Use and Head Injuries — Theory-based Methods

9.3.1 Learning outcomes

9.3.2 Terminology review

9.3.3 Winter sports helmet use and head injury

R Instructions

Hypothesis test

Use statistical analysis methods to draw inferences from the data

Confidence Interval

9.3.4 Effect of sample size

9.3.5 Take-home messages

9.3.6 Additional notes

9.4 Module 8 and 9 Lab: Poisonous Mushrooms

9.4.1 Learning outcomes

9.4.2 Poisonous Mushrooms

R Instructions

MODULE 10

Unit 2 Review

The following section contains both a list of key topics covered in Unit 2 as well as Module Review Worksheets.

10.0.1 Key Topics

Review the key topics for Unit 2 to review prior to the exams. All of these topics will be covered in Modules 6–9.

10.0.2 Module Review

The following worksheets review each of the modules. These worksheets will be completed during Melinda's Study Sessions each week. Solutions will be posted on Canvas in the Unit 2 Review folder after the study sessions.

10.1 Key Topics Exam 2

Descriptive statistics and study design

Hypothesis testing

Confidence interval

10.2 Module 6 Review - Simulation Methods - One Mean

10.3 Module 7 Review - Theory-based Methods - One mean

10.4 Module 7 and 8 Review

10.5 Module 8 and 9 Review

10.6 Group Exam 2 Review

MODULE 11

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

Placeholder

11.1 Vocabulary Review and Key Topics

11.1.1 Key topics

11.1.2 Vocabulary

Plotting a quantitative response with independent groups

Hypotheses

Simulation-based inference for a difference in means

Theory-based inference for a difference in means

11.2 Video Notes: Inference for Independent Samples

11.2.1 Course Videos

Theory-based method - Video 19.3TheoryTests

Optional Notes: Video Example (Video 19.3TheoryTests)

Confidence Interval - Video 19.3TheoryIntervals

Optional Notes: Video Example (Video 19.3TheoryIntervals)

Optional Notes: Simulation Testing for a Difference in Means: Video 19.1

Hypothesis Testing

Simulation-based method

Confidence interval

Optional Notes: Simulation Confidence Interval for a Difference in Means - Video 19.2

11.2.2 Concept Check

11.3 Activity 17: Does behavior impact performance?

11.3.1 Learning outcomes

11.3.2 Terminology review

11.3.3 Behavior and Performance

R instructions

Quantitative variables review

Ask a research question

Numerically Summarize the data

Use statistical inferential methods to draw inferences from the data

Hypothesis test

Notes on the null distribution

Confidence interval

Notes on the bootstrap distribution

11.3.4 Take-home messages

11.3.5 Additional notes

11.4 Activity 18: Moon Phases and Virtual Reality

11.4.1 Learning outcomes

11.4.2 Terminology review

11.4.3 Moon Phases and Virtual Reality

R instructions

Use statistical inferential methods to draw inferences from the data

11.4.4 Take-home messages

11.4.5 Additional notes

11.5 Module 11 Lab: Dinosaurs

11.5.1 Learning outcomes

11.5.2 Dinosaurs

R instructions

Use statistical inferential methods to draw inferences from the data

Hypothesis test

Communicate the results and answer the research question

Confidence interval

MODULE 12

Exploratory Data Analysis and Inference for Two Quantitative Variables

Placeholder

12.1 Vocabulary Review and Key Topics

12.1.1 Key topics

12.1.2 Vocabulary

Plotting two quantitative variables

Sample statistics for two quantitative variables

Hypotheses

Simulation-based inference for two quantitative variables

Theory-based methods for two quantitative variables

12.2 Video Notes: Regression and Correlation

12.2.1 Course Videos

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

Type of plot

Correlation

Slope

Coefficient of Determination

Multivariable plots - Video Chapter7

12.2.2 Concept Check

Theoretical Testing for Slope - Video 21.4to21.5TheoryTests

Optional Notes: Video Example (Video 21.4TheoryTests)

Theoretical Confidence Interval for Slope - Video 21.4TheoryInterval

Optional Notes: Video Example (Video 21.4TheoryInterval)

Video Notes: Inference for Two Quantitative Variables

Hypothesis Testing - Video 21.1

Simulation-based method

Confidence interval - Video 21.3

Simulation-based method

12.2.3 Concept Check

12.3 Activity 19: Moneyball — Linear Regression

12.3.1 Learning outcomes

12.3.2 Terminology review

12.3.3 Moneyball

Notes on two quantitative variables

R Instructions

Slope

Residuals

- 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.

12.4.2 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.

12.4.3 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
`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.

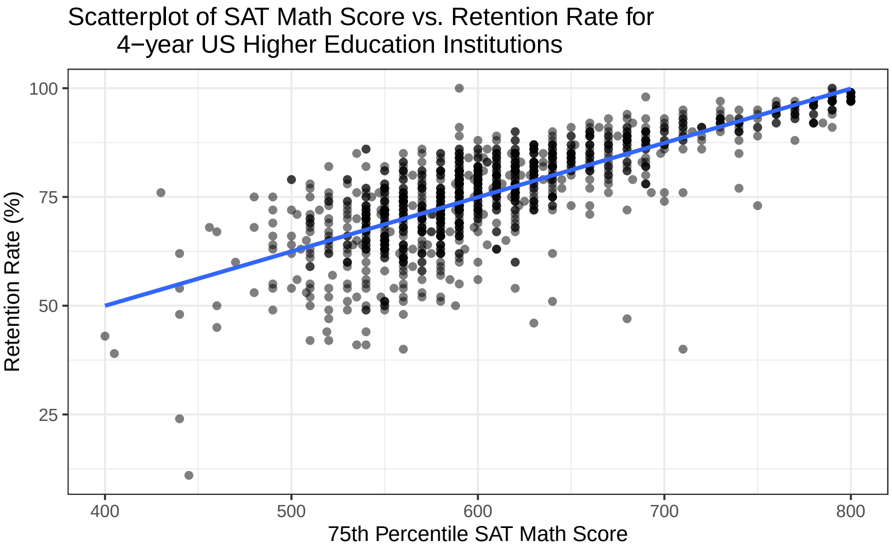
- Highlight and run lines 1–11 to load the data set and filter out the 2-year institutions.

```r
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.

```r
IPEDS %>% # Data sest 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, using the four characteristics of scatterplots, 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.

3. Interpret the value of slope.

4. Predict the retention rate for a 4-year US higher education institution with a 75th percentile SAT Math score of 440.

5. 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
```

6. 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: $r^2 = (r)^2$

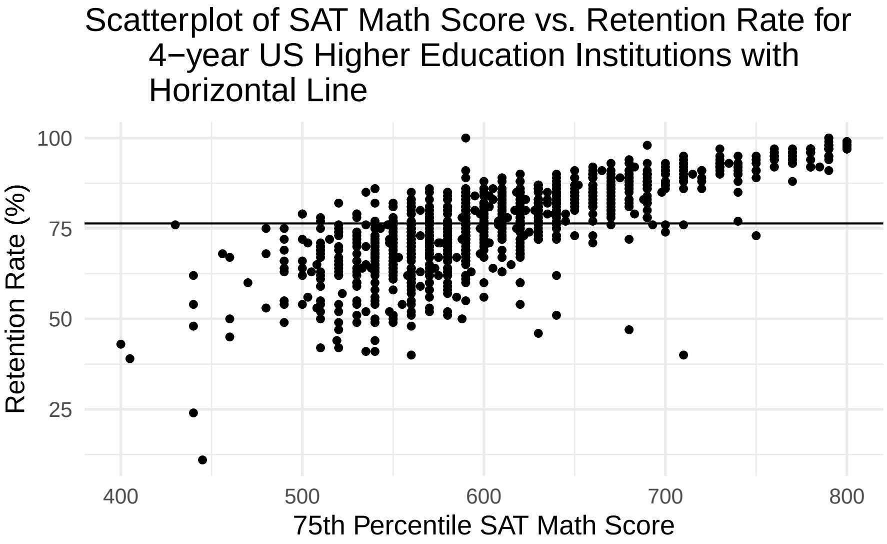
Use the variances of the response and the residuals: $r^2 = \dfrac{s_y^2 - s_{RES}^2}{s_y^2} = \dfrac{SST - SSE}{SST}$

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

The variance of the response variable, Retention (%), is $s^2_{Retention} = 138.386$ %^{2} and the variability in the residuals is $s^2_{RES} = 56.934$ %^{2}. Use these values to calculate the coefficient of determination.

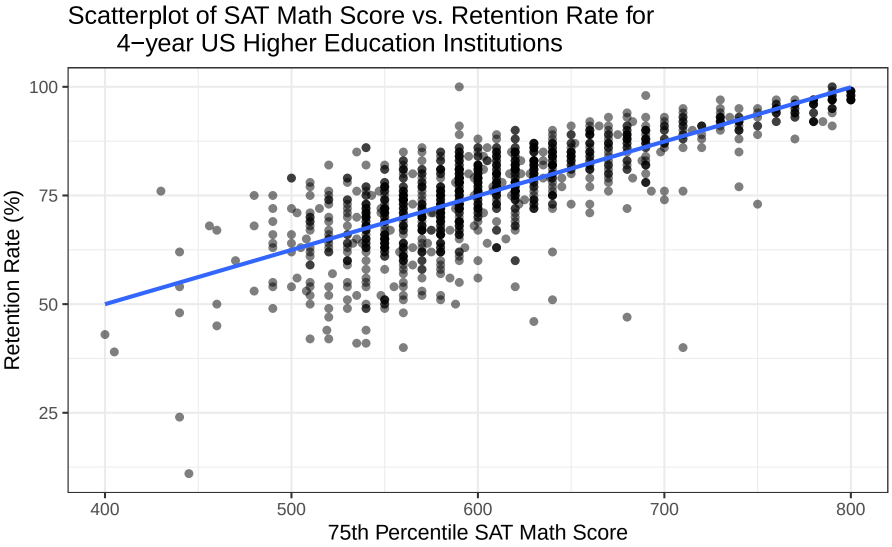
In the next part of the activity we will explore what the coefficient of determination measures.

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.

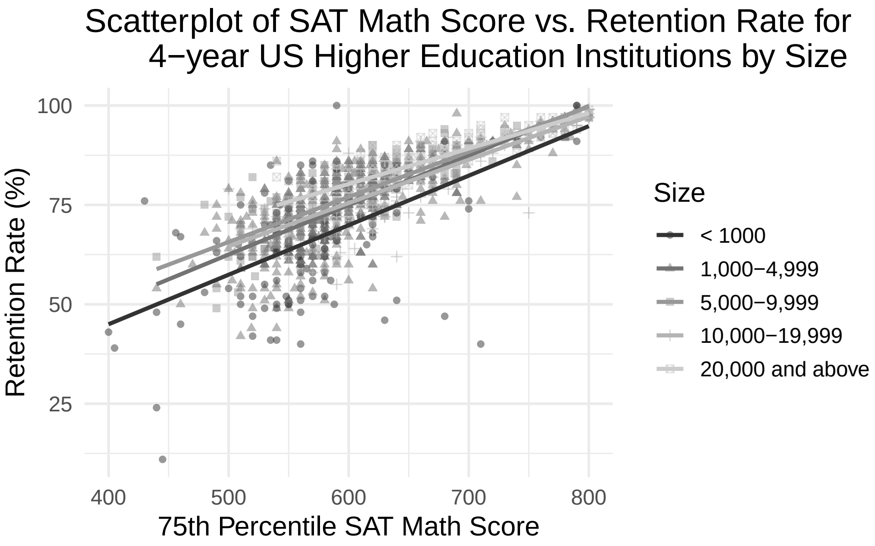
Calculate the value for r^2 using the values for SST and SSE provided below each of the previous graphs.

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

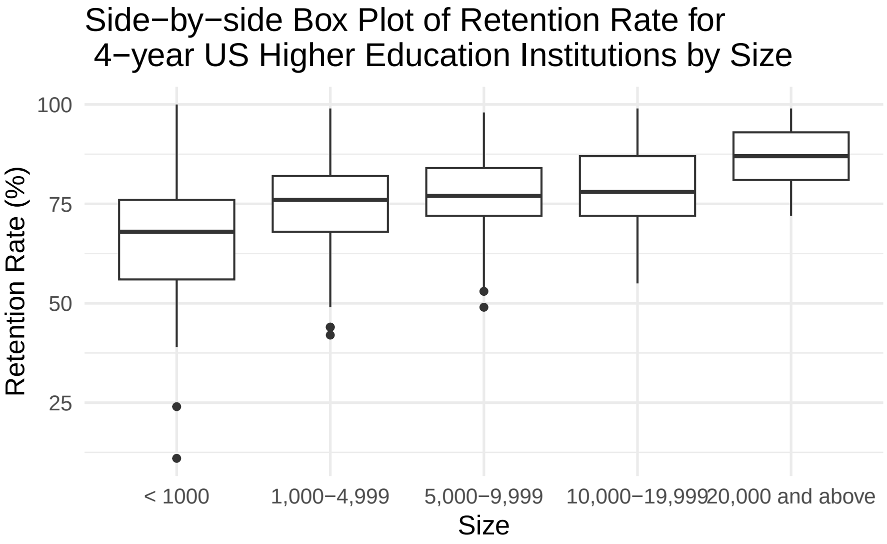
Multivariable plots

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 Size.

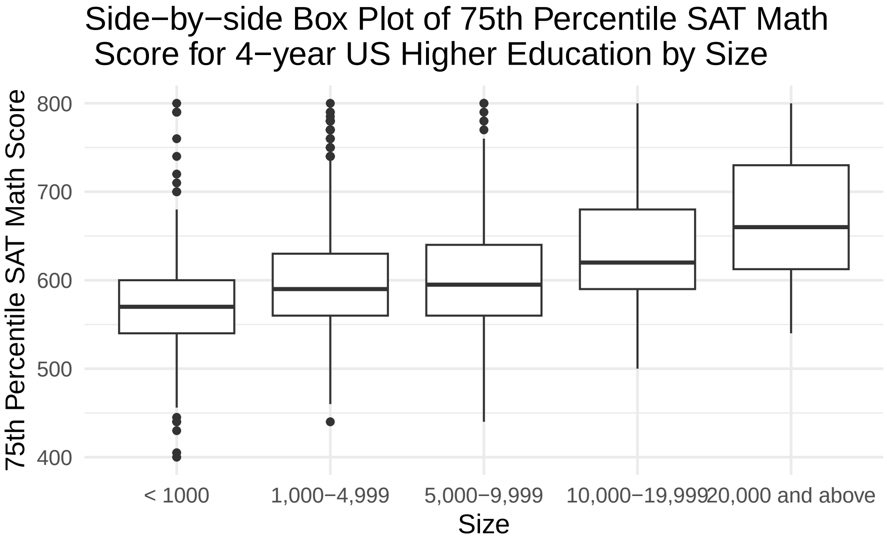
```
IPEDS$Size <- factor(IPEDS$Size, levels = c("< 1000", "1,000-4,999", "5,000-9,999",
                                            "10,000-19,999", "20,000 and above"))
IPEDS %>% # Data set pipes into...
    ggplot(aes(x = SATMath_75, y = Retention, shape = Size, color=Size))+  # Specify variables
    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 by Size") +
    # 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()
```



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



Side−by−side Box Plot of Retention Rate for 4−year US Higher Education Institutions by Size



Side−by−side Box Plot of 75th Percentile SAT Math Score for 4−year US Higher Education by Size

10. Is size of the higher education institution associated with retention rate? Is size of the higher education institution associated with 75th percentile SAT Math Score?

12.4.4 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**.

12.4.5 Additional notes

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

12.5 Activity 21: Golf Driving Distance

12.5.1 Learning outcomes

12.5.2 Terminology review

12.5.3 Golf driving distance

R Instructions

Plot review.

Conditions for the least squares line

Ask a research question

Summarize and visualize the data

Use statistical inferential methods to draw inferences from the data

Hypothesis test

Confidence interval

Communicate the results and answer the research question

Simulation-based hypothesis test

Simulation-based confidence interval

Multivariable plots

12.5.4 Take-home messages

12.5.5 Additional notes

12.6 Module 12 Lab: Big Mac Index

12.6.1 Learning outcomes

12.6.2 Big Mac Index

Summarize and visualize the data

Conditions for the least squares line

Ask a research question

Use statistical inferential methods to draw inferences from the data

Hypothesis test

Simulation-based confidence interval

Communicate the results and answer the research question

MODULE 13

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

Placeholder

13.1 Vocabulary Review and Key Topics

13.1.1 Key topics

13.1.2 Vocabulary

Simulation-based inference for a paired mean difference

Theory-based inference for a paired mean difference

13.2 Video Notes: Inference for Paired Data

13.2.1 Course Videos

Single categorical, single quantitative variables - Video Paired_Data

Paired vs. Independent Samples

Theory-based method - Video 18.3

t-distribution

Optional Notes: Video Example (Video 18.3)

Optional Notes: Simulation Inference for a Mean Difference - Video 18.1and18.2

Hypothesis testing

Simulation-based method

Confidence interval

Simulation-based method

13.2.2 Concept Check

13.3 Activity 22: Paired vs. Independent Samples

13.3.1 Learning outcomes

13.3.2 Terminology review

Notes on paired data

13.3.3 Paired vs. Independent Samples

13.3.4 Tattoo Effect on Sweat Rate

13.3.5 Exploring Paired Data

R Instructions

13.3.6 Take home messages

13.3.7 Additional notes

13.4 Activity 23: Color Interference

13.4.1 Learning outcomes

13.4.2 Terminology review

13.4.3 Color Interference

Identify the scenario

Ask a research question

MODULE 14

Unit 3 Review

The following section contains both a list of key topics covered in Unit 3 as well as Module Review Worksheets.

14.0.1 Key Topics

Review the key topics for Unit 3 to review prior to the exams. All of these topics will be covered in Modules 11–13.

14.0.2 Module Review

The following worksheets review each of the modules. These worksheets will be completed during Melinda's Study Sessions each week. Solutions will be posted on Canvas in the Unit 3 Review folder after the study sessions.

14.1 Key Topics Exam 3

Descriptive statistics and study design

Hypothesis testing

Confidence interval

14.2 Module 11 Review - Independent Samples

14.3 Module 12 Review - Regression

14.4 Module 13 Review - Paired Data

MODULE 15

Semester Review

Placeholder

15.1 Group Final Exam Review

15.2 Golden Ticket to Descriptive and Inferential Statistical Methods

In this course, we have covered descriptive (summary statistics and plots) and inferential (hypothesis tests and confidence intervals) methods for five different scenarios:

- one categorical response variable (Module 3 & 4)
- one quantitative response variable (Module 6 & 7) or paired differences in a quantitative variable (Module 13)
- two categorical variables (Module 8 & 9)
- one quantitative response variable and one categorical explanatory variable (Module 11)
- two quantitative variables (Module 12)

The "golden ticket" shown on the next page presents a visual summary of the similarities and differences across these five scenarios.

Scenario	One Categorical Response	One Quantitative Response or Paired Differences	Two Categorical Variables	Quant. Response and Categ. Explanatory (independent samples)	Two Quantitative Variables
Type of plot	Bar plot	Dotplot, histogram, boxplot	Segmented bar plot, Mosaic plot	Side-by-side boxplots, Stacked dotplots or histograms	Scatterplot
Summary measure	Proportion	Mean or Mean difference	Difference in proportions	Difference in means	Slope or correlation
Parameter notation	π	μ or μ_d	$\pi_1 - \pi_2$	$\mu_1 - \mu_2$	β_1 or ρ
Statistic notation	$\hat{p}$	$\bar{x}$ or $\bar{x}_d$	$\widehat{p_1} - \widehat{p_2}$	$\bar{x}_1 - \bar{x}_2$	b_1 or r
Null hypothesis	$H_0: \pi = \pi_0$	$H_0: \mu = \mu_0$ or $H_0: \mu_d = 0$	$H_0: \pi_1 - \pi_2 = 0$	$H_0: \mu_1 - \mu_2 = 0$	$H_0: \beta_1 = 0$ or $H_0: \rho = 0$
Conditions for simulation-based methods	Independent cases	Independent cases	Independent cases (within and between groups)	Independent cases (within and between groups)	Independent cases; Linear form
Simulation test (how to generate a null distn) p-value = proportion of null simulations at or beyond (H_A direction) the observed statistic	Spin spinner with probability equal to π_0, n times or draw with replacement n times from a deck of cards created to reflect π_0 as probability of success. Plot the proportion of successes. Repeat 10000 times. Centered at π_0	Shift the original data by adding $(\mu_0 - \bar{x})$ or $(0 - \bar{x}_d)$. Sample with replacement from the shifted data n times. Plot sample mean or sample mean difference. Repeat 10000 times. Centered at μ_0 for a single quantitative response or 0 for paired data.	Label cards with response values from original data; mix cards together; shuffle into two new groups of sizes n_1 and n_2. Plot difference in proportion of successes. Repeat 10000 times. Centered at 0.	Label cards with response variable values from original data; mix cards together; shuffle into two new groups of sizes n_1 and n_2. Plot difference in means. Repeat 10000 times. Centered at 0.	Separate the (x,y) pairs. Hold the x values constant; shuffle new y's to x's. Find the regression line for shuffled data; plot the slope or the correlation for the shuffled data. Repeat 10000 times. Centered at 0.
Bootstrap CI (how to generate a boot. distn) X% CI: $(\frac{1-X}{2}\%tile,$ $\left(X + \frac{1-X}{2}\right)\%tile)$	Label n cards with the original responses. Randomly draw with replacement n times. Plot the resampled proportion of successes. Repeat 10000 times. Centered at $\hat{p}$.	Label n cards with the original responses. Randomly draw with replacement n times. Plot the resampled mean difference. Repeat 10000 times. Centered at $\bar{x}$ for a single quantitative response or $\bar{x}_d$ for paired data.	Label n_1 cards with the original responses from group 1 and n_2 cards with the original responses from group 2. Keep groups separate. Randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Plot the resampled difference in proportion of successes. Repeat 10000 times. Centered at $\widehat{p_1} - \widehat{p_2}$	Label n_1 cards with the original responses from group 1 and n_2 cards with the original responses from group 2. Keep groups separate. Randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Plot the resampled difference in means. Repeat 10000 times. Centered at $\bar{x}_1 - \bar{x}_2$.	Label n cards with the original (explanatory, response) pairs. Randomly draw with replacement n times. Plot the resampled slope or correlation. Repeat 10000 times. Centered at b_1 for slope or r for correlation.
Theory-based distribution	Standard Normal	t- distribution with $n-1$ df	Standard Normal	t- distribution with min of $n_1 - 1$ or $n_2 - 1$ df	t- distribution with $n-2$ df
Conditions for theory-based hypothesis tests and confidence intervals	Independent cases; Number of successes and number of failures in the sample both at least 10.	Independent cases; $n < 30$ with no clear outliers OR $30 \le n < 100$ with no extreme outliers OR $n \ge 100$	Independence (within and between groups); Number of successes and number of failures in EACH sample all at least 10. (All four cell counts at least 10.)	Independent cases (within and between groups); In each sample, $n < 30$ with no clear outliers OR $30 \le n < 100$ with no extreme outliers OR $n \ge 100$	Linear form; Independent cases; Nearly normal residuals; Variability around the regression line is roughly constant.
Theory-based standardized statistic (test statistic)	$Z = \frac{\hat{p} - \pi_0}{SE_0(\hat{p})}$ $SE_0(\hat{p})$ $= \sqrt{\frac{\pi_0 \times (1-\pi_0)}{n}}$	$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})}$ OR $T = \frac{\bar{x}_d - 0}{SE(\bar{x}_d)}$ $SE(\bar{x}) = \frac{s}{\sqrt{n}}, SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}$	$Z = \frac{\widehat{p_1} - \widehat{p_2} - 0}{SE_0(\widehat{p_1} - \widehat{p_2})}$ $SE_0(\widehat{p_1} - \widehat{p_2})$ $= \sqrt{\widehat{p_{pool}} \times (1 - \widehat{p_{pool}}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE(\bar{x}_1 - \bar{x}_2)}$ $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$T = \frac{b_1 - 0}{SE(b_1)}$ $SE(b_1)$ is the reported standard error (std. error) of the slope term in the lm() output from R.
Theory-based confidence interval	$\hat{p} \pm z^* \times SE(\hat{p})$ $SE(\hat{p}) = \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}}$	$\bar{x} \pm t^* \times SE(\bar{x})$ $\bar{x}_d \pm t^* \times SE(\bar{x}_d)$ $SE(\bar{x}) = \frac{s}{\sqrt{n}}, SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}$	$\widehat{p_1} - \widehat{p_2} \pm z^* \times SE(\widehat{p_1} - \widehat{p_2})$	$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE(\bar{x}_1 - \bar{x}_2)$ $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$b_1 \pm t^* \times SE(b_1)$ $SE(b_1)$ is the reported standard error (std. error)

References

"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.

"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 Injuies 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., Dl. 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-Confimed 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 Heathy 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).