# Intro Stat Shiny

Chelsey Legacy

Department of Statistics

Iowa State University

Creative Component

Spring 2017

Committee Members: Amy G. Froelich, Major Professor

W. Robert Stephenson

Heike Hofmann

## Abstract

# Contents


**1 Introduction** . . . . . 4

**2 Literature Review** . . . . . 4

**3 Shiny Web Application Design** . . . . . 8

3.1 One Proportion: Sampling Distribution . . . . . 8

3.2 One Proportion: Confidence Intervals . . . . . 9

3.3 One Mean . . . . . 11

3.4 Two Proportions and Two Means . . . . . 12

3.5 Linear Regression: Correlation . . . . . 13

3.6 Linear Regression: Outliers . . . . . 14

3.7 Linear Regression: Equation . . . . . 15

3.8 ANOVA . . . . . 17

**4 Conclusions and Future Work** . . . . . 18

**5 References** . . . . . 18

5.1 Worksheets . . . . . 19

# 1 Introduction

Over the past 30 years, technology has become an integral part of teaching undergraduate statistics courses. Statistical software packages like JMP, StatCrunch, R, and Minitab are used to teach students the skills to conduct data analysis. Some of these same programs plus Java web-based applets have been used to teach students concepts in introductory statistics, particularly sampling distributions and inferential topics. Instructors select technology platforms based on a number of concerns particular to their audience and institution. One recent option, R's Shiny applications, provides an easy to use and interactive interface for introductory statistics students. The application can run on any web browser, allowing students and instructors to use the app on mobile devices, tablets, Chromebooks, and computers, with no cost to the students or instructor. As a relatively new option for instructional technology, there are few resources available for the introductory statistics course using the Shiny application in R. In this paper, we will outline the creation of an R shiny app and supporting supplemental worksheets to be used in conjunction with the app for the introductory statistics course. The Shiny app includes resources for teaching concepts in inference for a population proportion, population mean, difference between two population proportions, and the difference between two population means; descriptive linear regression and correlation, and ANOVA.

In section 2, we will discuss the literature related to technology in the classroom, specifically as it relates to teaching concepts in introductory statistics. In section 3, the development of the Shiny app is detailed, and we discuss the supplemental worksheets in Section 4. Conclusions and details of future work appear in Section 5.

# 2 Literature Review

Most undergraduate statistics courses make use of some sort of technology in the classroom from calculators to computer software programs. All the available technology has the ability to sim-

plify complex calculations in addition to making conceptual ideas more concrete. These are two main features of technology that has helped make it a staple tool for statistics education. Roy D. Pea provides an argument about why the computer as a cognitive technology is so important to the advancement in learning mathematics, in his article *Cognitive Technologies for Mathematics Education* (1987). Though his article focuses on the role of technology in math education there is a direct tie to statistics, in that he is discussing calculations and graphing data, both of which are essential elements of statistics courses. He introduces cognitive technologies as "any medium that helps transcend the limitations of the mind (e.g. attention to goals, short-term memory span), in thinking, learning, and problem-solving activities" (91). Computer applets and software are precisely the type of cognitive technologies being described that can help student thinking to go beyond just following formulas and returning answers.

Pea goes on to describe this idea further, "a common feature to all these cognitive technologies is that they make external the intermediate products of thinking (e.g. output of the component steps in solving a complex algebraic equation), which can then be analyzed, reflected upon, and discussed" (Pea, 1987, p. 91). With the help of computers, we are able to put thoughts and equations into visual and graphical representations that can then be viewed by many and discussed in an effort to clarify statistical concepts. The main goal of this project was to be used in helping statistics students gain some insight and confidence in working with complex concepts. Pea perfectly sums up this idea by saying that "the dynamic and interactive media provided by computer software make gaining an intuitive understanding (traditionally the province of the professional mathematician) of the interrelationships among graphic, equational, and pictorial representations more accessible to the software user."(1987, p. 96) Technologies such as R's Shiny web applications have the ability to provide insights to students through their interactivity and the output they provide that are not available through lectures or notes on a blackboard.

The use of these technologies in the curriculum had a profound impact on the pedagogy of statistics. Chance et al. provide a discussion in *The Role of Technology in Improving Student Learning of Statistics* about how courses are changing format. They specify that, "students are

evaluated less on their ability to manipulate formulas and look up critical values, and more on their ability to select appropriate tools (e.g., choosing techniques based on the variables involved), assess the validity of different techniques, utilize graphical tools for exploration of data, deal with messier data sets, provide appropriate interpretations of computer output, and evaluate and communicate the legitimacy of their conclusions“ (p. 3). They state, ”technology has expanded the range of graphical and visualization techniques to provide powerful new ways to assist students in exploring and analyzing data and thinking about statistical ideas, allowing them to focus on interpretation of results and understanding concepts rather than computational mechanics“ (p. 3). Though, technology has all the aforementioned benefits it is important that the technology should support the course learning objectives, and not restructure the objectives to pertain to learning the software. Chance et al, argue that technology should not be used ” merely for the sake of using technology”, but instead used for accessing, analyzing and interpreting large real data sets, automating calculations and processes, generating and modifying appropriate statistical graphics and models, performing simulations to illustrate abstract concepts and exploring 'what happens if...' type questions” (p.2-3).

There are increasingly many different software programs and applets available to supplement introductory statistics courses. Chance et al describe how ” the types of technology used in statistics and probability instruction can be broken into several categories: Statistical software packages, educational software, spreadsheets, applets/standalone applications, graphing calculators, multimedia materials, and data repositories.” (p.4) These options all have advantages and drawbacks highlighted in the article by Chance et al.

Choosing among these various options largely depends on the content of your course, your students expected competency on such programs, and funding available. Fathom, Minitab, and JMP are all point and click programs that allow users to enter and manipulate data, and they will provide graphical and summary statistics of the data. The drawback of such programs is that they cost the institution or the students money for downloads of the software. Of course R and R studio can also provide the graphical output and summary statistics needed by the students and they are

available for free. However, with both R and the point and click software there is a learning curve for the students. Many students may enter the course with little to no programming or even basic computer skills. To avoid the course becoming inaccessible to those students, the software chosen should be fairly straightforward and require no previous knowledge. Use of complex programs provides the problem of the students focusing so much on learning the software that they lose sight of the statistical concepts they are actually supposed to be learning.

In addition to software, there are many great web based java applets that are free to use and provide conceptual frameworks for simulations, p-values, and other intro stat concepts made available by Rossman and Chance (2004). These are helpful additions to already created lecture slides and materials, however they do not provide any supplemental materials that can help guide student interaction with the applets. As pointed out by Doi et al in *Web Application Teaching Tools for Statistics Using R and Shiny*, there are a multitude of available software and programs, however eventually an instructor may have difficulty finding an applet or software that perfectly suits their needs for their lectures (p. 1).

With monetary and time costs to consider, R‘s Shiny app technology is a great option for technology to incorporate into the classroom. R‘s package allows the developer to create an interactive app that is point and click based, and allows the user to perform complex tasks with the click of a button. In Doi et al’s article the authors provide a discussion on the advantages of using Shiny in the classroom. One of the main advantages is that the user never has to engage in writing or seeing the code if the developer chooses. The developer is the only one that must have a basic understanding of coding in R to get the app successfully running (p. 6, Doi et al, 2016). The authors explain that, ”all adjustments in the app are done by moving sliders or clicking buttons and corresponding updates to the output plot are virtually instantaneous. This leads to a much more fluid and dynamic presentation.” (p. 6). After creation and publication to the web the app can be run from a web browser so students can easily interact with it. Apps can be made for a variety of topics for the course and accessed from phones, tablets, and computers. This also allows the technology to be used by all students in classrooms other than just computer classrooms, which is

appealing to lectures without access to computer classrooms.

In addition, Doi et al discuss how the ease of design of a Shiny app allows it to be ”especially helpful when teaching concepts that are not in the standard curriculum or based on recent research” (p 6. 2016). Newer concepts may not already have applets or software available for classroom demonstrations, thus Shiny can be used to create something to fills that gap. It could also be used to create tailor made lecture materials for a specific course. Apps can be created based on the different topics in that course and thus creating a collection of relevant tools for a corresponding course.

# 3 Shiny Web Application Design

Designing and programming the R Shiny app was a large part of this project. There are many different concepts introduced to students in introductory statistics courses. A few of these concepts were chosen to be the focus of the Shiny app. The topics of focus are inference for one and two sample proportions, inference for one and two means, linear regression, and ANOVA. The Shiny app is composed of several tabs, each dedicated to one of these topics. Using the Shiny reference and examples pages a design was planned out for each of the different sections. Then, the applications were programmed for each section and subsection.

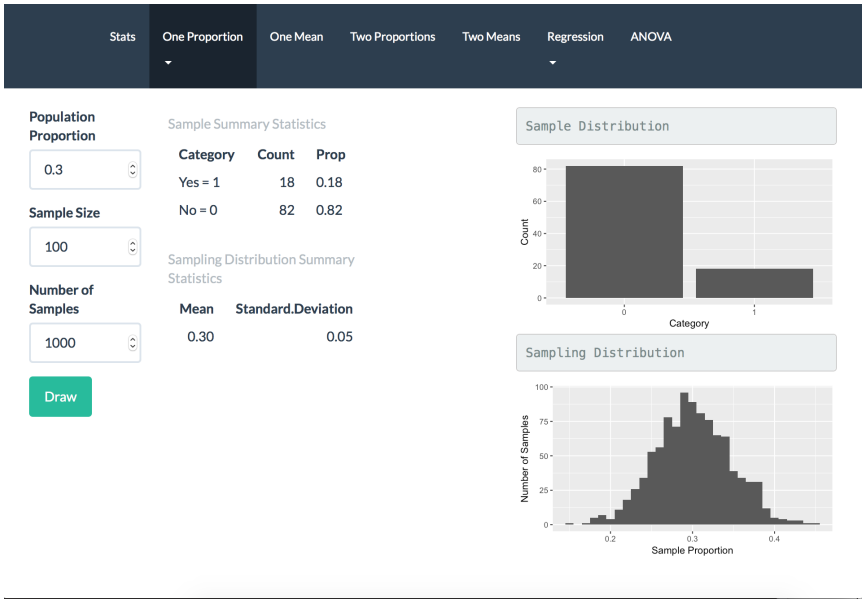
## 3.1 One Proportion: Sampling Distribution

As seen in Figure **??** the Shiny app opens up to the sampling distribution for one sample proportion. The design for this topic was based on several JMP scripts currently used in Statistics 101 at Iowa State University. This section is designed to help enhance student comprehension of the difference between a sample distribution and sampling distribution for a sample proportion. At the top of the application a variety of variables such as sample size, number of samples, and population proportion can be manipulated. Once these are set to the desired amounts the user can hit the

"Draw" button and the application will perform the simulation. The sample distribution graph is created using a binomial distribution with the proportion and sample size input by the user. If only one sample is created the Sampling Distribution graph will only display information from this one sample. If we choose to draw more than one sample, the Sample Distribution graph will display the last sample drawn. The Sampling Distribution will show a summary of all the proportions that were generated from the samples. The tables in the middle of the page provide summary information about the Sample and Sampling Distributions. The Sample Summary Statistics correspond to the Sampling Distribution, while the Summary Sampling Distribution Statistics correspond to the mean and standard deviation of the Sampling Distribution.

## 3.2 One Proportion: Confidence Intervals

The next section under the inference for one proportion is designed to demonstrate the effects of confidence level and sample size on the construction of confidence intervals. Figure **??** shows that for this section the number of samples, confidence level are available to manipulate as sliders. The confidence level allows the user to slide values from 80 to 99 percent confidence. The sample size allows for values from 25 to 500. The graph automatically displays 20 sample confidence intervals, created with sample proportions drawn from a binomial with $p = 0.5$, the input sample size, and $80\%$ confidence level. The lines represent the different confidence intervals and are colored according to whether or not they captured the true population parameter, 0.5. Light blue lines represent the confidence intervals that contained 0.5, while the orange lines represent those that did not contain 0.5.



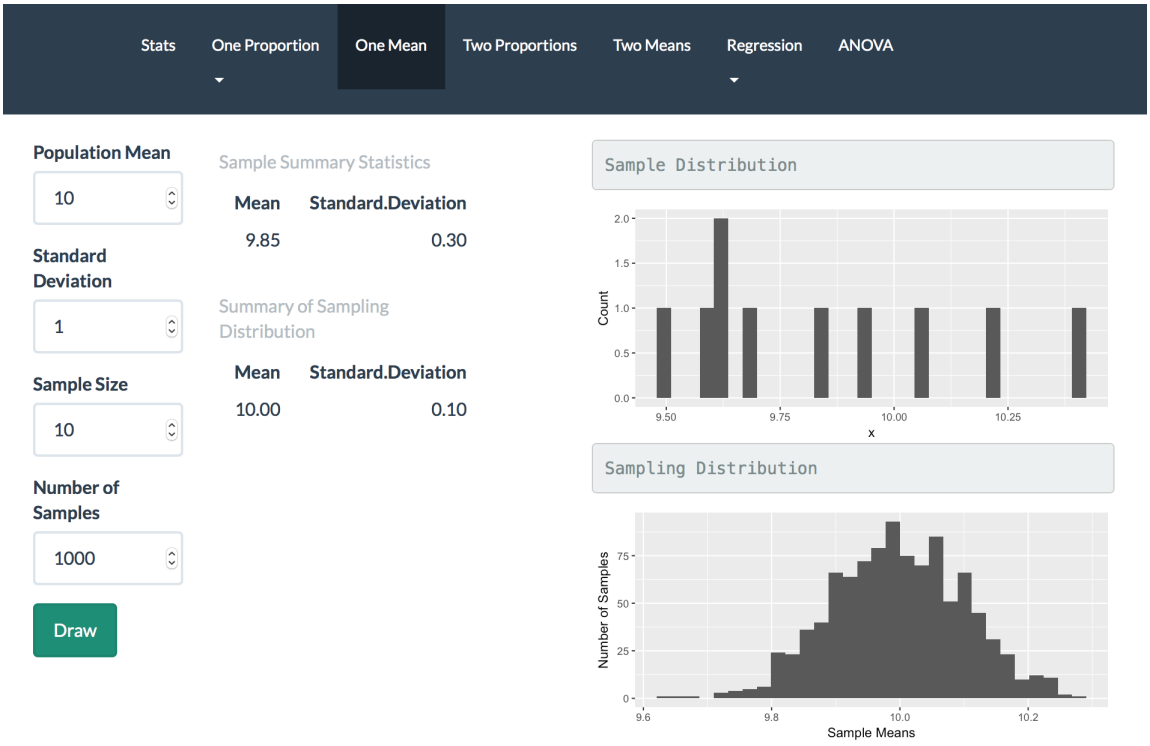

(a) Sampling Distribution Screen


(b) Confidence Interval Screen

Figure 1: Subsection of Inference for One Proportion.

## 3.3 One Mean

The section One Mean focuses on the concept of sampling distribution for one mean. They layout is similar to that of the inference for one proportion. The inputs allow the user to enter the mean, standard deviation, sample size, and number of samples. Once the ”Draw” button is clicked, the sampling distribution begins to appear with summary statistics in addition to the last drawn sample distribution and its summary statistics. The samples are drawn from a normal distribution with the mean and standard deviation input by the user. We can see a screen shot of this section in Figure **??**.



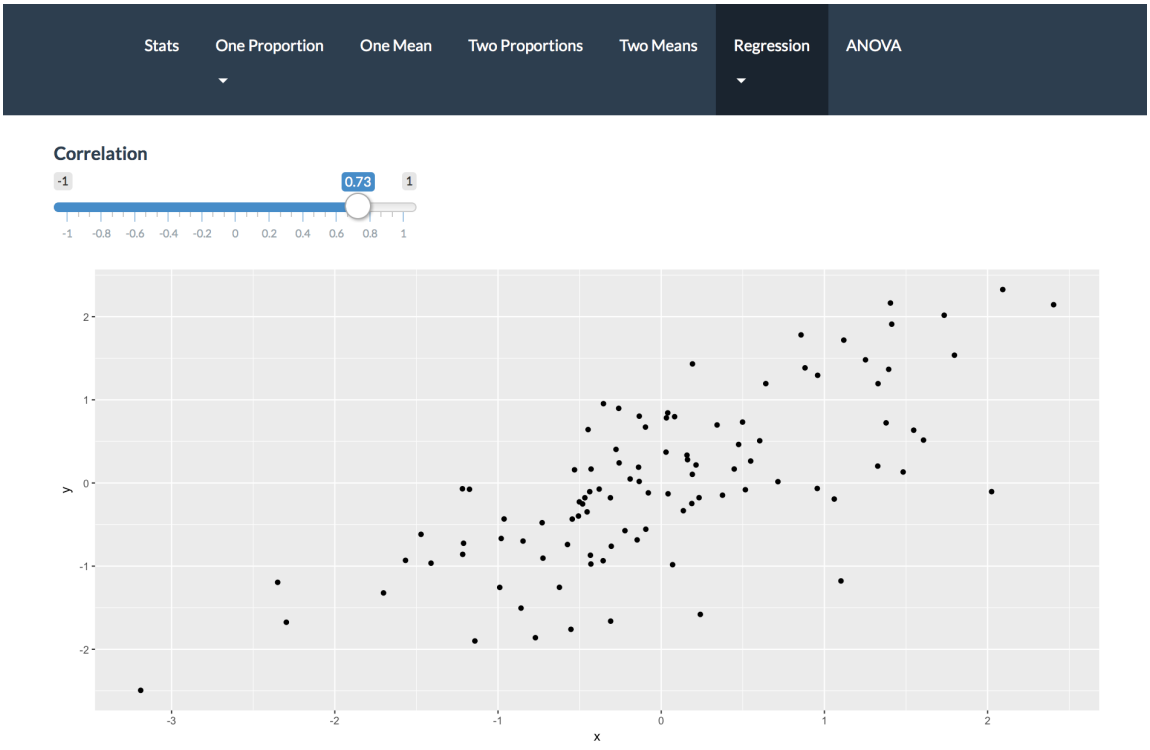
### 3.4 Two Proportions and Two Means

The sections for two proportions and two means are constructed similarly to the single proportion and single means sections, as seen in Figure **??**. They each aim to display a sampling distribution of sample statistic differences. The two proportions tab allows the user to input the population proportion for two different groups, the sample size for those groups, and the number of samples they wish to draw. It then draws from a binomial with the specified parameters for each group. The results of the sample are shown in the Sample Distribution graphs. The table of Sample Summary information displays the proportion for each group as well as the difference in those proportions. The Sample Distribution graph is created by repeating this process many times and saving the differences in the two proportions each time. The summary information for the Sampling Distribution is displayed in the Sampling Distribution Summary table.

The Two Means tab recreates the same process except with calculating the difference in means for two different groups. The user can input a specific mean and standard deviation for each of the two groups. The samples will then be generated form a normal distribution with those specifications. The setup of this section is similar to that of Two Proportions where we are given both the Sample Distribution and the Sampling Distribution, in addition to tables that provide the summary statistics for those graphs.


## 3.5 Linear Regression: Correlation

Another section featured in the Shiny application is based on linear regression topics. The correlation page features a slider that allows the user to select correlation values for a random plot of points. Figure **??** gives a a screen shot of this section of the Shiny app. This allows the user to see what different correlations look like ranging from -1 to 1. In order to make this plot, a multivariate normal random data frame is created that has 100 sets of data points designed with a specific sigma matrix. This sigma matrix changes based on the user input correlation value, but keeps 1 as the variance for both x and y. Thus, the XY coordinates are generated with a set correlation. These points are then plotted on the graph and update as the slider is moved.




## 3.6 Linear Regression: Outliers

The outliers tab under Linear regression is designed to display multiple data sets with varying outliers. The data sets are all identical except for the one outlier point, which will be set in a red color. Once a data set is selected the user can select to fit a line of least squares to the data, both with and without the outlier. Once the box is checked to fit a line to the data, the line is placed on the graph, blue for the line fit with the outlier and green slashes for a line fit to the data without the outlier. The output will also display the intercept, slope and coefficient of determination for the lines on the graph.

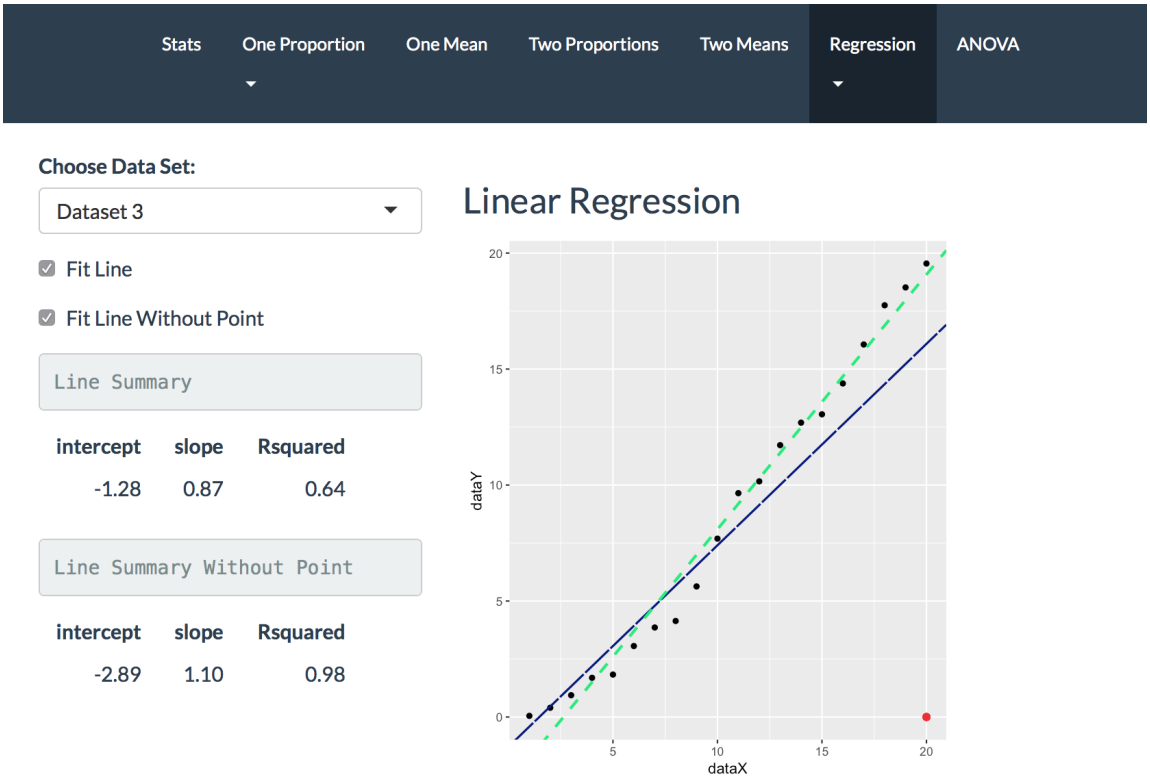



Figure 2: Outliers in Linear Regression

## 3.7 Linear Regression: Equation

The Equation section of the Shiny app breaks down the components of an equation and provides interpretations for intercepts and slopes. In addition, the user can manipulate the intercept and slope of the points on the graph in order to see how it changes the points and the the line fit to them. The data for this plot is generated using the same multivariate random normal data generating function in R as in the correlation section. This time, the input slope value is set as the correlation in the sigma matrix. The data is then adjusted to meet the intercept requirements by shifting the points

created with the specific slope to pass through the user input intercept value. The data on the plot will update as the intercept and slope sliders are moved. When the “Fit Line” box is selected the equation for that least squares line will be displayed. It will have the same intercept and slope as the sliders display. The user can then select the ”Intercept” box to see a red dot appear at the y-intercept on the plot. They will also be shown an intercept interpretation that is specific the the intercept they see in the plot. Similarly, when the ”Slope” box is checked there is an interpretation of the slope that is output for the user. The graph will display the rise and run of the slope to help illustrate the interpretation for the slope. A horizontal green line represents the one unit increase in the explanatory variable, while a vertical blue line indicates the increase or decrease in the response variable as a result of the increase in the explanatory variable. The user can manipulate the sliders and the check boxes as they see fit, and the plot and output will continually update. Figure **??** displays the screen for Equation when all boxes are checked.

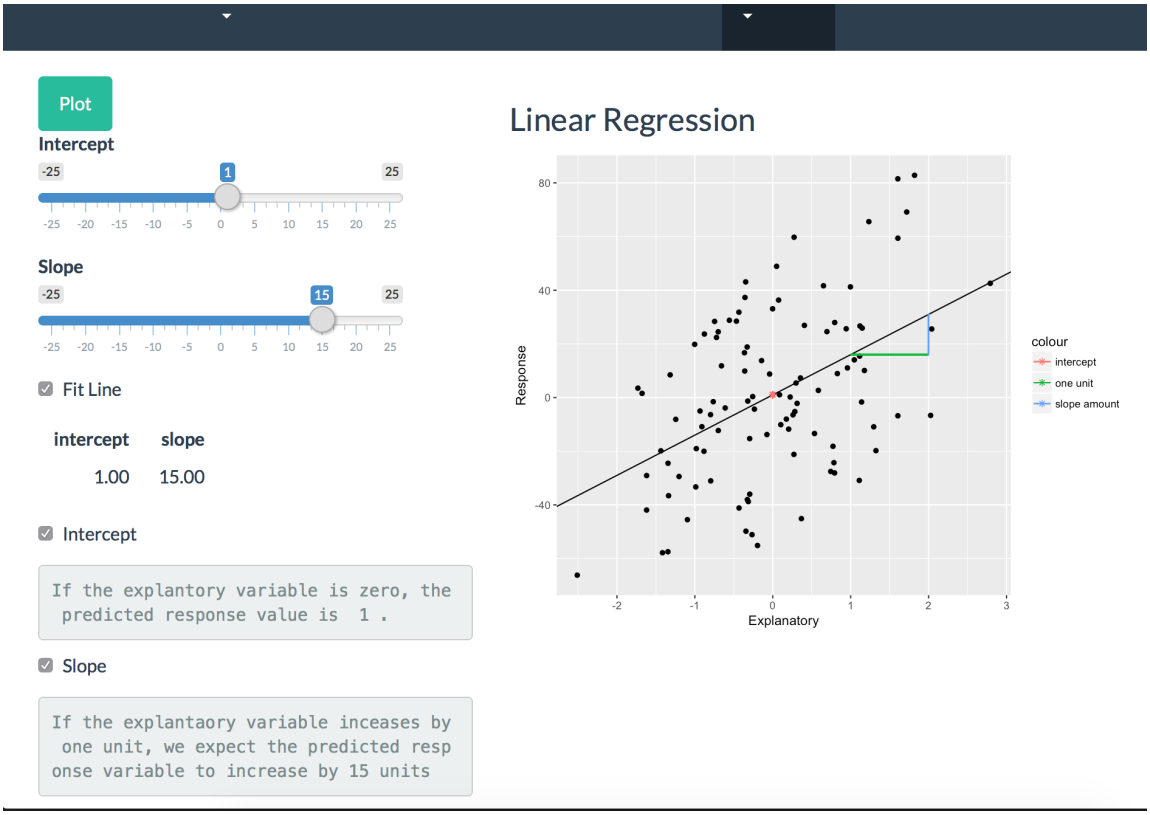
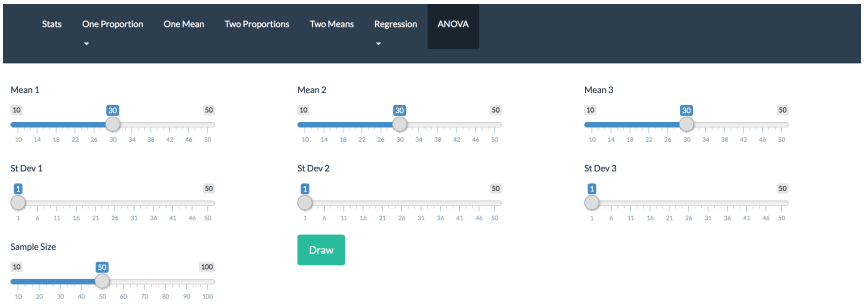


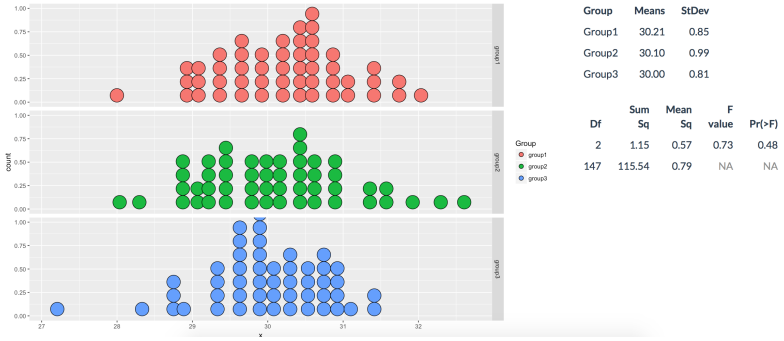

Figure 3: Equation Tab

## 3.8 ANOVA

The last tab of the shiny application is the ANOVA tab. This tab explores the relationship among the means of 3 different groups. The user can input the means, standard deviations, and sample sizes for each of the three groups. This can be seen in Figure **??**. Then R will generate the samples for each group using a normal distribution with the input parameters for each group. Then, the graph will display a dotplot for each of those groups. Next to the graph, is an ANOVA table for the differences in the means along with the summary information for each of the three groups. The output for this part of the Shiny app can be seen in Figure **??**.




(a) ANOVA Section top




| Group | Means | StDev |
|---|---|---|
| Group1 | 30.21 | 0.85 |
| Group2 | 30.10 | 0.99 |
| Group3 | 30.00 | 0.81 |

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Group | 2 | 1.15 | 0.57 | 0.73 | 0.48 |
| | 147 | 115.54 | 0.79 | NA | NA |

(b) ANOVA Section lower

Figure 4: ANOVA Sections

# 4 Conclusions and Future Work

# 5 References

# Bibliography

[1] Chance, B., Ben-Zvi, D.,& Garfield, J., Medina, E. (2007). The Role of Technology in Improving Student Learning of Statistics.

[2] Doi, J., Potter G., Wong J., Alcaraz I., & Chi, P. (2016). Web Application Teaching Tools for Statistics Using R and Shiny.

[3] Roy D. Pea. Cognitive Technologies for Mathematics Education. A. Schoenfeld. Cognitive science and mathematics education. Hillsdale, pp.89-122, 1987. <hal-001990547>

[4] Rossman, A.& Chance, B. (2004), The *Rossman/Chance Applet Collection*. Available at www.rossmanchance.com/applets.

## 5.1 Worksheets

Another component of this project was creating worksheets that correspond to each of the sections of the Shiny app. Though the app can stand alone to help vizualize various topics, the worksheets walk the user through using each section and try to help illustrate ideas. The worksheets for each section work to help demonstrate the following ideas:

- One Proportion and One Mean: Leads user through drawing one sample and seeing how it is displayed in the Sample Distribution vs in the Sampling Distribution. Then, leads to drawing more samples and seeing that the Sampling Distribution ends up centered around population

proportion.

- Confidence Interval: Walk the user through fixing the confidence level and seeing what happens to the width of the intervals as the sample size in increased from 25 to 500. In the second question the sample size is fixed at 250 and the confidence level is increased from 80 to 99%. Through manipulation of these two components the user is quickly able to see the relationship between confidence interval width, the sample size, and the confidence level.

- Two Proportions and Two Means: takes the user the student through taking multiple samples to build the sampling distribution which ends up being approximately normal, similar to the one proportion and one mean inference sections.

- Corrleation: takes the user through observing the structures of the data as the slider moves from a correlation of -1 to 1. They will take note of the direction and form of the data as the correlation changes.

- Outliers: leads the user through the linear model selection process of fitting data both with and without an outlier and determining which model best fits the data. Users will also look at multiple outliers and determine whether they have high leverage, high residual or both.

- Equation: guides the user through manipulating the slope and intercept on the plot and porviding interpretations of those values.

- ANOVA: gives the user a small introduction to the ANOVA F test followed by some questions asking them to explore the relationship between the mean, standard deviation, sample size and the resulting p-value and conclusion to the F-test.

# QUESTIONS FOR SHINY APP

CHELSEY LEGACY

## 1. Sampling Distribution for a Proportion

1. Take 1 sample of size 100 with a population proportion of 0.3. Draw a graph of the sample distribution, and write the proportion of those chosen that fell into the "yes" category.

Note that the sampling distribution has one single bar at the value for the "yes" proportion.

2. Now take 100 samples of size 100 with a population proportion 0.3. What do you notice about the shape of the sampling distribution now?

3. Now take 1000 samples of size 100 with a population proportion of 0.3. How would you describe this sampling distribution? What are the mean of this distribution?

## 2. Confidence Interval

1. Draw samples at confidence level 95%, increasing the sample size from 25 to 500 slowly. What do you notice about the length of the intervals as the sample size increased? Why?

2. Now increase the confidence level from 80% to 99% slowly, keeping the sample size at 250. What do you notice about the length of the intervals as the confidence level increased? Why?

## 3. Inference for one mean

1. Take 1 sample of size 25 with a population mean of 20 and a standard deviation of 3. Draw a graph of the sample distribution, and write the mean of your sample.

Note that the sampling distribution has one single bar at the value of the mean for your sample.

2. Take 100 samples of size 25 with a population mean of 20 and a standard deviation of 3. What do you notice about the sampling distribution now?

3. Take 1000 samples of size 25 with a population mean of 20 and a standard deviation of 3. How would you describe this sampling distribution? What is the mean for this sampling distribution?

## 4. Inference for two Proportions

1. Set the proportion for group 1 to 0.3 and the proportion for group 2 to 0.4. Also, set the number of samples to 1, and draw a sample. What $\hat{p}$ for each sample?

2. What is your sample statistic for the difference in proportions from question 1?

Note: The sampling distribution has one sample at your sample difference in proportions value.

3. Set number of samples to 100 and redraw. Describe the shape, center, and symmetry of your sampling distribution.

## 5. Inference for two Means

1. Draw 1 sample with sample size 50 for each group. Set group 1 mean to be 20 and group 2 mean to be 21. Let the standard deviations for both be 1. What are your sample means for each group?

2. What is your sample statistic for the difference in means from question 1?

Note: The sampling distribution has one sample at your sample difference in means value.

3. Set number of samples to 100 and redraw. Describe the shape, center, and symmetry of your sampling distribution.

## 6. Correlation

1. Set the correlation to -0.9. Describe the form, strength, and direction of the points.
2. Set the correlation to -0.5. Describe the form, strength, and direction of the points.
3. Set the correlation to 0.5. Describe the form, strength, and direction of the points.
4. Set the correlation to 0.9. Describe the form, strength, and direction of the points.
5. What do you notice about the form and strength as you increased the correlation from -0.9 to 0.9?

## 7. Outliers

1. Select "Dataset 1", and observe the point in red. Click "Fit Line" to get a least squares line to the data. Write the equation of the line below.

2. Next, remove the red point and refit the least squares line by checking the "Fit Line Without Point" box. Write the equation for that line below.

3. Do you think that fitting the data without the red point improved the fit of the line to the data? Why?

4. Select "Dataset 2", and observe the point in red. Click "Fit Line" to get a least squares line to the data. Write the equation of the line below.

5. Next, remove the red point and refit the least squares line by checking the "Fit Line Without Point" box. Write the equation for that line below.

6. Do you think that fitting the data without the red point improved the fit of the line to the data? Why?

7. Choose "Dataset 3" from the dropdown. Fit the least squares regression line both with and without the red outlier. Do you think the point has high leverage? Do you think the point has a large residual? Explain.

High leverage:

Large residual:

8. Choose "Dataset 4" from the dropdown. Fit the least squares regression line both with and without the red outlier. Do you think the point has high leverage? Do you think the point has a large residual? Explain.

High leverage:

Large residual:

## 8. Regression

1. Set the intercept to -10 and the slope to 5. Click "Fit Line" and write the equation for the least squares regression line.

2. Write the y-intercept interpretation. Then, check the "Intercept" box and see if your interpretation is correct. Note, the red dot on the plot inidcation the location of the y-intercept.

3. Write the slope interpretation. Then, check the "Slope" box and see if your interpretation is correct.

4. Set the intercept to 20 and the slope to -15. Click "Fit Line" and write the equation for the least squares regression line.

5. Write the y-intercept interpretation. Then, check the "Intercept" box and see if your interpretation is correct.

6. Write the slope interpretation. Then, check the "Slope" box and see if your interpretation is correct. What is the difference between this interpretation and the one in number 3? Explain this difference.

## 9. ANOVA

Analysis of variance, or ANOVA, is used to compare means among groups. An ANOVA table can provide a test to see if there is a significant difference among means for multiple groups. If the p-value for the test is significant based on a set alpha level, we know that at least one group mean is significantly different than the other group means.

1. Set all means to 30 and all standard deviations to 1. State the p-value and conclusion for a test of difference in means.

2. Set the means to 20, 30, and 30 respectively. Set all standard deviations to 5. Draw a sample and state the p-value and conclusion.

3. Set the mean for groups 1 and 2 to 30, and mean of group 3 to 32. Set all standard deviations to 5. Take a sample with sample size 100 and note what the p-value is for the ANOVA test. Write a conclusion based on your p-value.

4. Next, keep the settings the same as above, but draw a sample of size 10. Note the p-value, and write a conclusion based on this p-value.

5. What generalization can be made regarding sample size and test outcome after observing what happened in the previous questions?

## 10. Pre-Questions for Students in 101

1. Explain what it means to have 90% confidence in a confidence interval.
2. As we increase the confidence level of a confidence interval what happens to the width of the interval?

## 11. Post-Questions for Students in 101

1. Explain what it means to have 90% confidence in a confidence interval.
2. As we increase the confidence level of a confidence interval what happens to the width of the interval?
3. Comments on the Shiny app for confidence level:
4. What is the purpose of an ANOVA test?
5. Comments on the Shiny app for ANOVA tests:

## 12. Pre-Questions for Students in 400 Level Stats

1. What is the y-intercept of a line in words?
2. What is the slope of a line in words?

## 13. Post-Questions for Students in 400 Level Stats

1. What is the y-intercept interpretation for a line with an intercept of 2?
2. What is the interpretation of the slope for a line with a slope of -3?
3. Comments of the Shiny app section for linear regression.
4. Explain the difference between a sample distribution and a sampling distribution.
5. Comments on the Shiny app for a sampling distribution for means.