Using the R Library RPanel for GUI-Based Simulations in Introductory Statistics Courses

A Senior Project
Presented to
The Faculty of the Statistics Department
California Polytechnic State University, San Luis Obispo

In Partial Fulfillment
Of the Requirements for the Degree
Bachelor of Science

By
Ryan M. Allison

Senior Project Advisor:
Soma Roy, Ph.D.

May, 2012
© 2012 Ryan M. Allison
# Table of Contents

1. Introduction .......................................................................................................................... 5

2. Executive Summary ............................................................................................................... 6

3. How to Download and Install R ........................................................................................... 7

   3.1 Installation – Windows .................................................................................................. 9

   3.2 Installation – Mac OS X ............................................................................................. 15

   3.3 Installation – Linux .................................................................................................... 23

4. Two Way Tables Applet ....................................................................................................... 29

   4.1 Instructor Guide ......................................................................................................... 29

       4.1.1 Applet Introduction ............................................................................................ 29

       4.1.2 Simulation .......................................................................................................... 33

       4.1.3 Normal Approximation ..................................................................................... 36

       4.1.4 Exact Test .......................................................................................................... 38

   4.2 Student Guide .............................................................................................................. 39

       4.2.1 Applet Introduction ............................................................................................ 39

       4.2.2 Simulation .......................................................................................................... 39

       4.2.3 Normal Approximation ..................................................................................... 40

       4.2.4 Exact Test .......................................................................................................... 41

5. Two by Three Tables Applet ............................................................................................... 42

   5.1 Instructor Guide ......................................................................................................... 42

       5.1.1 Applet Introduction ............................................................................................ 42

       5.1.2 Sum of Pairwise |Differences| ........................................................................... 45

       5.1.3 Sum of |Phati-Phat| ............................................................................................ 48

       5.1.4 Chi-Square Statistic .......................................................................................... 50

       5.1.5 Chi-Square Test ................................................................................................. 53

   5.2 Student Guide .............................................................................................................. 55

       5.2.1 Applet Introduction ............................................................................................ 55

       5.2.2 Sum of Pairwise |Differences| ........................................................................... 55

       5.2.3 Sum of |Phati-Phat| ............................................................................................ 56

       5.2.4 Chi-Square Simulation ..................................................................................... 56

       5.2.5 Chi-Square Test ................................................................................................. 57

6. r x k Tables Applet .............................................................................................................. 58
6.1 Instructor Guide .......................................................................................................................... 58
  6.1.1 Applet Introduction.................................................................................................................. 58
  6.1.2 Reading In Data ...................................................................................................................... 59
  6.1.3 Sum of Pairwise |Differences| .............................................................................................. 59
  6.1.4 Sum of |Phati-Phat| ................................................................................................................. 64
  6.1.5 Chi-Square Statistic ................................................................................................................ 68
  6.1.6 Chi-Square Test ...................................................................................................................... 71

6.2 Student Guide .............................................................................................................................. 73
  6.2.1 Applet Introduction.................................................................................................................. 73
  6.2.2 Sum of Pairwise |Differences| .............................................................................................. 73
  6.2.3 Sum of |Phati-Phat| ................................................................................................................. 74
  6.2.4 Chi-Square Simulation ........................................................................................................... 74
  6.2.5 Chi-Square Test ...................................................................................................................... 75

7. Analysis of Variance Applet.......................................................................................................... 76
  7.1 Instructor Guide .......................................................................................................................... 76
    7.1.1 Applet Introduction.................................................................................................................. 76
    7.1.2 Reading In Data ...................................................................................................................... 77
    7.1.3 Descriptive Statistics ............................................................................................................ 77
    7.1.4 Simulate No Association ....................................................................................................... 79
    7.1.5 ANOVA / F Test .................................................................................................................... 84

  7.2 Student Guide .............................................................................................................................. 86
    7.2.1 Applet Introduction.................................................................................................................. 86
    7.2.2 Simulate No Association ....................................................................................................... 87
    7.2.3 ANOVA / F Test .................................................................................................................... 87

8. Simple Linear Regression Applet .................................................................................................... 88
  8.1 Instructor Guide .......................................................................................................................... 88
    8.1.1 Applet Introduction.................................................................................................................. 88
    8.1.2 Reading In Data ...................................................................................................................... 89
    8.1.3 Descriptive Statistics ............................................................................................................ 89
    8.1.4 Simulate No Association ....................................................................................................... 90
    8.1.5 Simple Linear Regression ..................................................................................................... 95

  8.2 Student Guide .............................................................................................................................. 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2.1</td>
<td>Applet Introduction</td>
<td>97</td>
</tr>
<tr>
<td>8.2.2</td>
<td>Simulate No Association</td>
<td>98</td>
</tr>
<tr>
<td>8.2.3</td>
<td>Simple Linear Regression</td>
<td>98</td>
</tr>
<tr>
<td>9.</td>
<td>Feedback from Students</td>
<td>99</td>
</tr>
<tr>
<td>10.</td>
<td>Next Steps</td>
<td>100</td>
</tr>
</tbody>
</table>
1. Introduction

I believe that the R Project (http://www.r-project.org/), a free and open-source statistics package, is rather underutilized in introductory statistics courses as R has powerful graphical capabilities that would allow students to visualize data and the effects of simulating no association. Introductory statistics courses are missing a large opportunity to expose students to a powerful package that many professionals use in industry. However, due to the standard user interface to the program being nearly entirely code-driven unlike other frequently used packages like Minitab, JMP, or web-based Java applets, utilizing the software in an introductory course would require students either learn how to code or be forced to copy and paste pre-existing code. In my opinion, both of these methods abstract the important statistical concepts to be learned by the students due to the inaccessibility of the program to introductory students. Frequently, students learn the code rather than learning the relevant statistical methods.

However, the R library RPanel (http://cran.r-project.org/web/packages/rpanel/) allows a programmer to build a Graphical User Interface (GUI) that runs R code in the background without the user having to know any actual R code. I believed that utilizing this library to develop various applets would make a great way to introduce students to the R package and allow students to run various interactive simulations. Also, R would allow for simulating statistical concepts that did not previously have corresponding applets. In addition, these allow students to utilize their own data sets to analyze and simulate no association. As such, I developed five different simulation applets that utilize the RPanel library: a 2x2 table simulation, a 2x3 table simulation, an rxk table simulation, a regression simulation, and a one-way ANOVA simulation.

This paper documents these applets, provides guides for both instructors and students to follow with suggested questions to check for understanding, and documents the results of a survey distributed to students after utilizing one of these applets in an introductory statistics course.

To gain access to the applets, please email Ryan Allison at: ryanmallison@gmail.com.
2. Executive Summary

As a student, I noticed that the statistical package R (http://www.r-project.org) would have several benefits of its usage in the classroom. One benefit to the package is its free and open-source nature. This would be a great benefit for instructors and students alike since it would be of no cost to use, unlike other statistical packages. Due to this, students could continue using the program after their statistical courses and into their professional careers. It would be good to expose students while they are in school to a tool that professionals use in industry. R also has powerful graphical abilities that would allow students to visualize their data and the effects of simulating no association on their data. Utilizing R would also allow students to read in their own data in order to further explore analyses and see more examples rather than just built in datasets as with other applets. Finally, simulation-based instruction that R is capable of does not necessarily need a formal test-statistic unlike other methods of instruction. This would allow students to use more intuition in learning statistical concepts.

Unfortunately, there are several challenges to using R in the classroom. Primarily, the user interface is designed around coding. This abstracts the important statistical concepts that students are trying to learn. Many times, students try to learn the code rather than learn the relevant statistics and end up not learning from the technology. This was the key motivation for my project. The idea was that I would build a graphical user interface (GUI) on top of R to allow students to explore statistical concepts and be exposed to R.

As such, I created five applets to allow students to explore simulation and distribution based tests found in introductory statistics courses:

1. 2x2 Table
2. 2x3 Table
3. rxk Table
4. Regression
5. One-Way Analysis of Variance

These applets have the benefit over other similar approaches in that they allow students to see the animation of the effect of each repetition on simulation of no association, students do not need to have any knowledge of code or programming to utilize the applets, and most importantly, students and instructors have the ability to use their own data sets.

After utilizing several of these applets in introductory statistics courses, students responded very positively. Specifically, a survey was distributed to students following use of the ANOVA applet on a lab and on a homework assignment. The survey indicated that the project was a resounding success. Further, all students recommended use of the applets in future statistics courses. Specifically, students felt that it was helpful in furthering their knowledge of ANOVA. Students felt that it was particularly helpful because of the step-by-step animation showed students the simulation as it was happening. It allowed students to understand what was occurring in a much more visual fashion. Some students even preferred the applets over other packages like Minitab that students were more familiar with due to the straightforward and visual nature. Finally, students also recommended that statistics professors use the applet in future courses, as well.
3. How to Download and Install R

As R is an open-source software, it is free to download and install at http://www.r-project.org

First, click on the CRAN (Comprehensive R Archive Network) to proceed to the download section as in Figure 1. Then, pick the download server mirror that is located geographically closest to your location as in Figure 2.

Next, as in Figure 3, pick your platform you would like to install R on (Windows, Mac OS X, Linux).
Then, click the large download link to download and install R as in Figure 4.

![Download R 2.15.0 for Windows](image)

Figure 4: Download the R Package

If you are on a Macintosh platform, make sure to install a separate component found in the tools link.

![Subdirectories](image)

Download the tcltk file in Figure 5 and install it, as well as R.

![tcltk-8.5.5-x11.dmg](image)

Figure 5: Download tcltk
3.1 Installation – Windows
Installing on a PC is simple. After you have downloaded the R installer, double click on the .exe file. This will open the following window (Figure 6):

![Figure 6: Choose Language](image)

Click OK to choose your desired language, and then click next to proceed through the installation in Figure 7.

![Figure 7: Click Next](image)
Click Next to accept the GNU General Public License agreement (Figure 8).

![Figure 8: Accept the GNU GPL](image)

Click Next to choose the default installation path as in Figure 9.

![Figure 9: Choose Installation Path](image)
Click Next to accept the default components for installation (Figure 10).

![Figure 10: Choose Default Components](image)

Accept the default startup options by clicking Next.

![Figure 11: Accept Default Startup Options](image)
Click Next to accept the default Start Menu Folder (Figure 12).

![Select Start Menu Folder](image1.png)

**Figure 12: Choose Start Menu Folder**

Proceed to installation by clicking the Next button.

![Select Additional Tasks](image2.png)

**Figure 13: Proceed to Installation**
Once you have hit next, the installation will proceed as in Figure 14.

![Figure 14: R Installation](image1)

Once this has completed, the page from Figure 15 will appear.

![Figure 15: Setup Completion](image2)
Click Finish to complete the installation. You may now launch the R environment as in Figure 16 and follow the further guides.

Figure 16: R Environment
3.2 Installation – Mac OS X

Installing on a Macintosh is quite straightforward. After you download the R and tcltk files, double click on the R installer file. This will open the window found in Figure 17:

![R OSX Installation](image1)

You will now see a description of the R components like in Figure 18. Press “Continue” to proceed:

![Description of the R Components](image2)
Click the continue button again after viewing the GNU General Public License Agreement:

Figure 19: GNU GPL

Press continue and agree as in Figure 20 to accept the GNU General Public License software license agreement:

Figure 20: Agree to GNU GPL
As in Figure 21, you may choose your install location. It is suggested to simply click the Install button and accept the default values.

Now, R will proceed to install to your computer as in Figure 22.
Following this, the installer will run until it completes and shows the screen found in Figure 23:

![Completed Installation](image)

**Figure 23: Completed Installation**

Now, unlike the Windows and Linux versions, Macintosh platforms must install the tcltk package, as well. Double click that installer to begin a similar process. The disk image will mount and present a window, as in Figure 24:

![tcltk Installer File](image)

**Figure 24: tcltk Installer File**
From here, double click the tcltk.pkg icon to launch the installer found in Figure 25.

![Figure 25: tcltk Installer](image)

Following that, you will be shown the following description of the tcl/tk package as in Figure 26. Click the continue button to proceed through the installer.

![Figure 26: tcl/tk Description](image)
You will be shown the software license agreement description as in Figure 27. Press “Continue” to proceed through the installation process.

Click the continue button to bring up the software license agreement acceptance page found in Figure 28. Click Agree to accept the GNU General Public License.
Click Continue again to accept the default installation location as in Figure 29. (In this case, the main hard drive.)

![Figure 29: Installation Location](image)

Now, click the Install button as in Figure 30 to begin the installation process.

![Figure 30: Install tcl/tk](image)
Once the installer completes, you will be shown the screen found in Figure 31. Now, you will be able to click the Close button.

Now, the R installation is complete, and you may launch the R environment found in Figure 32. From here, you may follow the standard R guides.
3.3 Installation – Linux

There is a fair amount of variation in the method of installation depending on the type of Linux distribution you wish to install on. As such, it is recommended to follow the current installation guides located at: http://cran.stat.ucla.edu/bin/linux/.

For reference, the following is a description of the installation process on Debian-based Linux distributions (Debian, Ubuntu, Linux Mint, etc.) shown in the GNOME desktop environment in Debian.

First, locate the type of distribution you would like to install on (in this case, Debian) as in Figure 33:

![Index of /bin/linux](image)

**Figure 33: Choose Distribution**
Next, find the repository line that refers to the R packages as in Figure 34:

![Figure 34: R Repository](image)

Then, open the Software Sources dialog (Figure 35) located in the System menu:

![Figure 35: Software Sources](image)
Click on the Third-Party Software tab (Figure 36) and click on the add button to paste in the repository link:

![Figure 36: Add R Repository](image)

The package manager will prompt you to update your sources to include the new repository as in Figure 37:

![Figure 37: Refresh Sources](image)
Once you click the reload button, the software manager will update the sources like in Figure 38:

![Figure 38: Repository Updating](image)

Once this completes, we can begin installation of R. In this case, I am using the APT package manager in the command line. Install the packages r-base and r-base-dev in a similar fashion to Figure 39. If you receive an error about the public key not being available for the R repository, you may ignore it.

![Figure 39: Install R](image)
Accept the download and installation of the software as in Figure 40:

Figure 40: Accept Download/Installation

If it asks if you would like to accept the packages without verification, just accept as in Figure 41:

Figure 41: Accept Installation
The package manager will be installing R like in Figure 42:

![Figure 42: R Installation](image)

Once this completes, just launch R by typing `R` at the command line and you will see a similar page to Figure 43:

![Figure 43: Launch R](image)

From here, R works just like the other platforms.
4. Two Way Tables Applet

4.1 Instructor Guide

4.1.1 Applet Introduction

This applet will allow students to explore categorical data through simulation. In order to prepare the applet, students simply double click on the Two Way.RData workspace, which launches the R environment on computers with R installed and loads the appropriate functions. To launch the applet, students simply need to launch the function by typing “twoway()”. The parentheses are necessary as this instructs the R environment to launch the function named “twoway”. Interested students, however, may view the code by launching the function within R without the parentheses.

Once the function is launched, the function will automatically download and load the required packages. Once this completes (which should take no more than a few seconds), the panel from Figure 44 will open:

![Figure 44: Two-Way Panel](image.png)
Students can press the Instructions button to see, as found in Figure 45, the instructions for using the applet:

![Figure 45: Two-Way Instructions](image)

Students are required to begin by thinking about how the data should be analyzed by clicking on the appropriate Explanatory and Response variables. If students click an option that does not make sense in the context of the data, students are presented with an error message that helps guide their thinking.

For example, if they choose a variable to be either both explanatory and response (Figure 46):

![Figure 46: Choosing Incorrect Explanatory / Response](image)
Or perhaps they choose the incorrect order (Figure 47):

![Image of incorrect order]

**Figure 47: Incorrect Order**

Once they choose the appropriate combination (in this case, Explanatory: Drug, Response: Abstain/Not Abstain), three buttons, as in Figure 48, will appear allowing students to continue with exploration of these data:

![Image of complete panel available]

**Figure 48: Complete Panel Available**
However, students can also enter in their own data to explore simply by changing the values in the table. After changing the Explanatory/Response labels, though, students must press the Refresh Panel button to update the applet. For example, if we change to a different dataset, the applet appears as found in Figure 49.
4.1.2 Simulation

Beginning with the simulation button, students will see the following simulation panel appear (Figure 50):

![Simulation Panel](image)

Figure 50: Simulation Panel

Students are able to choose which statistic they would like to simulate via the radio buttons: Difference in Proportions, Odds Ratio, and Relative Risk.
Upon the first shuffle of difference in proportions, the following graphics device will appear as in Figure 51:

![Graphics Device](image)

*Figure 51: Graphics Device*

For any number of shuffles less than 50, students will be able to watch the simulated data be animated for each additional repetition. That is, the simulated data bar graph will adjust for each additional shuffle, and the histogram of the sum of pairwise $|differences|$ will change for each shuffle. The simulation pauses for one half of a second in between shuffles. Conversely, for any number of shuffles greater than 50, the simulation will simply output the final simulation rather than animating the process for efficiency.
As such, an example of 1000 shuffles would appear similarly to Figure 52:

![Figure 52: 1000 Shuffles](image)

Students may also click the checkbox “Show the Last Randomized Table,” which will output the last simulated table to the console to see the upper right bar chart numerical representation and see how it compares to that of the original data. An example would be (Output 1):

<table>
<thead>
<tr>
<th></th>
<th>Abstain</th>
<th>Not</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nicotine</td>
<td>65</td>
<td>399</td>
</tr>
<tr>
<td>Placebo</td>
<td>66</td>
<td>392</td>
</tr>
</tbody>
</table>

**Output 1: Last Randomized Table**
In addition, students can also choose odds ratio (Figure 54) and relative risk (Figure 55) as statistics to simulate:

![Standard Deviation](image1)

**Figure 53: Odds Ratio**

![Standard Deviation](image2)

**Figure 54: Relative Risk**

### 4.1.3 Normal Approximation

If students click on the Normal Approximation button, students will see the panel found in Figure 55:

![Normal Approximation](image3)

**Figure 55: Normal Approximation**
If students run a 2 Sample Z-Test, students will see the following output (Output 2):

<table>
<thead>
<tr>
<th>Nicotine</th>
<th>Placebo</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.18750000</td>
<td>0.09606987</td>
<td>0.09143013</td>
</tr>
</tbody>
</table>

*Output 2: 2 Sample Z-Test*

Similarly, students can run a Chi-Square Test on the data as in Output 3:

```r
Pearson's Chi-squared test with Yates' continuity correction
data:  datamatrix
X-squared = 15.0657, df = 1, p-value = 0.0001038
```

*Output 3: Chi-Square Test*

Students may also ask for confidence intervals of various statistics at specified confidence levels. For example, for Difference in Proportions in Output 4:

```r
$`95 % Confidence Interval`
Lower Bound Upper Bound
 0.04682491  0.13603536

$`P-value`
  Abstain
  5.882848e-05
```

*Output 4: Difference in Proportions*

For relative risk (Output 5):

```r
[1] "Relative Risk"
$`95 % Confidence Interval`
Lower Bound Upper Bound
 1.907099    1.996310

$`P-value`
  Abstain
  0
```

*Output 5: Relative Risk*

And for odds ratio (Output 6):

```r
[1] "Odds Ratio"
$`95 % Confidence Interval`
Lower Bound Upper Bound
 5.946679   12.933547

$`P-value`
  Abstain
  6.331942e-28
```

*Output 6: Odds Ratio*
4.1.4 Exact Test

Finally, students can click the Exact button and obtain the Fisher’s Exact Test p-value as in Output 7:

| Fisher's Exact Test P-Value | 6.957584e-05 |

Output 7: Fisher’s Exact Test

Some suggested questions for students are:

- How does your simulated data change with each “shuffle” of your observed data? How does this compare to your observed data?
- How would your analysis change if you have twice the number people in each category, but the proportion stays the same? How will your p-value change?
4.2 Student Guide

4.2.1 Applet Introduction
This applet will allow you to explore categorical data through simulation. In order to prepare the applet, double click on the “Two Way.RData” file. To launch the applet, simply launch the function by typing “twoway()” as in Figure 56. It may take a few seconds to launch, as it may have to download and install a few files before it opens.

When you launch the function, the following panel will open (Figure 57):

If you would like any clarification on how to use the applet, press the Instructions button at any time.

This particular data file allows you to look at the relationship between nicotine and a placebo and its effects on abstaining from smoking, but you may change the data values however you like to explore other data. If you change the Explanatory/Response labels, make sure to press the “Refresh Panel” button to update the applet.

Start by considering which of the variables will affect the other. That is, which variable will be the explanatory and which will be the response. Once you explore the correct combination, you will be able to proceed with exploring the relationship.

4.2.2 Simulation
Start by pressing the simulation button to open the Simulation Panel found in Figure 58. From here, you are able to choose which statistic you would like to simulate via the radio buttons: Difference in Proportions, Odds Ratio, and Relative Risk.
Once you shuffle the data for the first time, take note of the simulated version of the data. How do those barcharts compare to the barcharts in our original data? Does the relationship between the two groups change?

Continue by changing the number of shuffles to 10. (You have to press the “Enter” key when changing any values to have the applet accept the change.) From the animation, how does the relationship between the two groups change? Is there a large change in the barcharts with each shuffle?

Add another 40 shuffles to get to 50 total. Take note of the histogram of your observed statistic. Is the histogram symmetric? What value is it centered around? Do these make sense with our data?

Add another 950 shuffles and take a look at the final histogram similar to Figure 59. Try changing the observed value of that statistic to the value for the original data and click the “Refresh Graph” button. How extreme is that value? What does that tell you about the association in the context of our data?

You may also click the checkbox “Show the Last Randomized Table,” which will show the last simulated table to see the upper right bar chart expressed as a table. How does it compare to that of the original data?

### 4.2.3 Normal Approximation

If you click on the “Normal Approximation” button, students will see a similar panel to the previous simulation panel. From this panel, you can run two different tests based on a normal approximation, as well as obtain confidence intervals based on the three statistics from above. What do these tell you in the context of our data? Are you arriving at similar conclusions? Does that make sense?

Examples of output are:

#### 2 Sample Z-Test

<table>
<thead>
<tr>
<th>Nicotine</th>
<th>Placebo</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.18750000</td>
<td>0.09606987</td>
<td>0.09143013</td>
</tr>
</tbody>
</table>

Output 8: 2 Sample Z-Test
**Chi-Square Test**

Pearson's Chi-squared test with Yates' continuity correction

```
data:  datamatrix
X-squared = 15.0657, df = 1, p-value = 0.0001038
```

**Difference in Proportions**

```
[1] "Difference in Proportions"
$'95 % Confidence Interval'
Lower Bound Upper Bound
0.04682491 0.13603536

$'P-value'
Abstain
5.882848e-05
```

**Relative Risk**

```
[1] "Relative Risk"
$'95 % Confidence Interval'
Lower Bound Upper Bound
1.907099 1.996310

$'P-value'
Abstain
0
```

**Odds Ratio**

```
[1] "Odds Ratio"
$'95 % Confidence Interval'
Lower Bound Upper Bound
5.946679 12.933547
```

**4.2.4 Exact Test**

Finally, you can click the "Exact" button and obtain p-value from Fisher’s Exact Test. Does this follow your previous conclusions? What does that tell you about the association in the context of our data?

**Exact Test Output**

```
Fisher's Exact Test P-Value
6.957584e-05
```

If you wish to further explore the data, consider how your analysis would change if you have twice the number people in each category, but the proportion stays the same. How will your p-value change?
5. Two by Three Tables Applet

5.1 Instructor Guide

5.1.1 Applet Introduction

This applet will allow students to explore categorical data through simulation. In order to prepare the applet, students simply double click on the Twoby Three.RData workspace, which launches the R environment on computers with R installed and loads the appropriate functions. To launch the applet, students simply need to launch the function by typing “twobythree()”. The parentheses are necessary as this instructs the R environment to launch the function named “twobythree.” Interested students, however, may view the code by launching the function within R without the parentheses.

Once the function is launched, the function will automatically download and load the required packages. Once this completes (which should take no more than a few seconds), the panel in Figure 60 will open:

![Two-by-Three Applet Panel](image)

Figure 60: Two-by-Three Applet
Students can press the Instructions button to see the instructions for the applet as found in Figure 61:

![Instructions](image)

**Figure 61: Instructions**

Students are required to begin by thinking about how the data should be analyzed by clicking on the appropriate Explanatory and Response variables. If students click an option that does not make sense in the context of the data, students are presented with an error message that helps guide their thinking.

For example, if they choose a variable to be either both explanatory and response as in Figure 62:

![Choosing Incorrect Explanatory/Response](image)

**Figure 62: Choosing Incorrect Explanatory/Response**
Or perhaps students choose the incorrect order as in Figure 63:

![Figure 63: Incorrect Explanatory / Response Order](image)

Once they choose the appropriate combination (in this case, Explanatory: Donor Decision, Response: Default), four buttons will appear allowing students to continue with exploration of these data (Figure 64):

![Figure 64: Complete Two-by-Three Applet](image)
However, students can also enter in their own data to explore simply by changing the values in the table. After changing the Explanatory/Response labels, though, students must press the Refresh Panel button to update the applet. For example, if we change to a different dataset, the applet will look similar to Figure 65:

![Figure 65: Changed Data](image)

**5.1.2 Sum of Pairwise |Differences|**

Beginning with the simulation of the sum of pairwise |differences| for the original dataset, students will see the following simulation panel appear (Figure 66):

![Figure 66: Sum of Pairwise |Differences| Panel](image)
Upon the first shuffle, the following graphics device will appear similar to Figure 67:

For any number of shuffles less than 50, students will be able to watch the simulated data be animated for each additional repetition. That is, the simulated data bar graph will adjust for each additional shuffle, and the histogram of the sum of pairwise \(|\text{differences}|\) will change for each shuffle. The simulation pauses for one half of a second in between shuffles. Conversely, for any number of shuffles greater than 50, the simulation will simply output the final simulation rather than animating the process for efficiency.
As such, an example of 1000 shuffles would appear similar to Figure 68:

In addition, information about the pairwise differences of the last shuffled data and the original data is displayed in the R console:

```
[1] "Original Data"
   Opt-In Opt-Out Neutral Sum of Absolute Differences
Pairwise Differences  -9    32    32                      73
[1] "Last Simulated Table"
   Opt-In Opt-Out Neutral Sum of Absolute Differences
Pairwise Differences  -23   -20   -12                      55
```

Output 14: Simulated vs. Original Data
5.1.3 Sum of \( |\text{Phati-Phat}| \)

If students click on the Sum of \( |\text{Phati-Phat}| \) button, students will see a similar simulation panel to the previous simulation as in Figure 69:

![Figure 69: Sum of \( |\text{Phati-Phat}| \)](image)

Similar to the previous simulation, students can see how the last shuffled data compares to the original data in the R console as in Output 15:

```
[1] "Original Data"
   Opt-In Opt-Out Neutral Sum of \( |\text{Phati-Phat}| \)
P-Hats 0.4181818  0.82   0.7857143     0.516725
[1] "Last Simulated Table"
   Opt-In Opt-Out Neutral Sum of \( |\text{Phati-Phat}| \)
P-Hats 0.6363636  0.6   0.7678571     0.202301
```

Output 15: Original vs. Simulated \( |\text{Phati-Phat}| \)
In addition, the animation feature is also present. Just as with the previous simulation, students can enter in the observed value of the statistic from the original data to see how the shuffled data compares (represented by the red line) similar to Figure 70:

Figure 70: Entering the Observed Value
5.1.4 Chi-Square Statistic

The chi-square statistic simulation has a few extra features to enable students to explore the distribution similar to Figure 71:

The bottom left histogram attempts to plot a normal distribution with mean and standard deviation based on the shuffled data to show students that a normal distribution would not be appropriate for these data. The bottom right graph plots a more appropriate chi-square distribution over the data, while allowing students to explore the distribution through degrees of freedom. That is, students can increase and decrease the degrees of freedom through buttons on the simulation panel (Figure 72):
If we incorrectly increase the degrees of freedom to 10 and press the Refresh Graph button, we notice that the fit of the curve is much worse than the previous fit:
However, if students correctly decrease the degrees of freedom to 2 and press the Refresh Graph button, students can see that the distribution is much more appropriate for these data as in Figure 74:

![Figure 74: Correct Degrees of Freedom](image)

Finally, if students press the Show Last Randomized Table button, students can see what the last shuffled table is that the bar graph is representing and see how it compares to that of the original data similarly to Output 16:

<table>
<thead>
<tr>
<th></th>
<th>Opt-In</th>
<th>Opt-Out</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donor</td>
<td>34</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>Not</td>
<td>21</td>
<td>14</td>
<td>18</td>
</tr>
</tbody>
</table>

|                  | Simulated Distribution | 24.284 | 2 | 0 |

Output 16: Last Shuffled Table
5.1.5 Chi-Square Test

Finally, after exploring different simulations, students can explore the results of a chi-square test. If students click the Chi-Square Test button, students can see a shaded graph of the distribution and p-value similar to Figure 75:

Relevant information about the data will also be outputted to the R console including the Observed Data, Expected Counts, and Chi-Square Cell Contributions as in Output 17:

```
$`Observed Data`
   Opt-In  Opt-Out Neutral
Donor   23     41     44
Not     32      9      12

$`Expected Data`
   Opt-In  Opt-Out Neutral
Donor  36.89441 33.54037 37.56522
Not   18.10559 16.45963 18.43478

$`Chi-Square Cell Contributions`
   Opt-In  Opt-Out Neutral
Donor  5.232625 1.659076 1.102254
Not   10.662708 3.380759 2.246103
```

Output 17: Chi-Square Test Information
Some suggested questions for students are:

- How does your simulated data change with each “shuffle” of your observed data? How does this compare to your observed data?
- Would a normal approximation be appropriate for these data? Why or why not?
- What happens to the chi-square distribution as you increase the degrees of freedom? As you decrease the degrees of freedom? Which provide a better “fit” to your shuffled data? What value fits your shuffled data best? Why do you think this is?
- What observed value of your statistic would be appropriate to use for these data?
- How would your analysis change if you have twice the number people in each category, but the proportion stays the same? How will your p-value change?
5.2 Student Guide

5.2.1 Applet Introduction

This applet will allow you to explore categorical data through simulation. In order to prepare the applet, double click on the “Two By Three.RData” file. To launch the applet, simply launch the function by typing “twobythree()” as in Figure 76. It may take a few seconds to launch, as it may have to download and install a few files before it opens.

When you launch the function, the following panel will open (Figure 77):

If you would like any clarification on how to use the applet, press the Instructions button at any time.

This particular data file allows you to look at the relationship between default status and donor decision, but you may change the data values however you like to explore other data. If you change the Explanatory/Response labels, make sure to press the “Refresh Panel” button to update the applet.

Start by considering which of the variables will affect the other. That is, which variable will be the explanatory and which will be the response. Once you explore the correct combination, you will be able to proceed with exploring the relationship.

5.2.2 Sum of Pairwise |Differences|

Start by pressing the Sum of Pairwise |Differences| button. From here, press the “Shuffle” button to start shuffling your data.

Once you shuffle the data for the first time, take note of the simulated version of the data. How do those barcharts compare to the barcharts in our original data? Does the relationship between the two groups change?
Continue by changing the number of shuffles to 10. (You have to press the “Enter” key when changing any values to have the applet accept the change.) From the animation, how does the relationship between the two groups change? Is there a large change in the barcharts with each shuffle?

Add another 40 shuffles to get to 50 total. Take note of the histogram of your observed statistic. Is the histogram symmetric? What value appears the most? Do these make sense with our data?

Add another 950 shuffles and take a look at the final histogram. Try changing the observed value of that statistic to the value for the original data and click the “Refresh Graph” button. How extreme is that value? What does that tell you about the association in the context of our data?

You may also click the checkbox “Show the Last Randomized Table,” which will show the last simulated table to see the upper right bar chart expressed as a table. How does it compare to that of the original data?

5.2.3 Sum of |Phati-Phat|

If you click on the “Sum of |Phati-Phat|” button, you will see a similar panel to the previous simulation panel. From this panel, you can run a similar test to the previous one. Go ahead and repeat the same procedure from above. How does this statistic differ? Does it provide similar results to the ones you obtained above?

5.2.4 Chi-Square Simulation

Now moving on to the “Chi-Square Simulation” panel, you should see the panel appear as in Figure 79. Feel free to follow the same procedure as before. Does the normal distribution in the lower left corner appear appropriate for these data? Once you run 1000 repetitions, try adjusting the degrees of freedom of the green chi-square distribution curve in the lower right panel by pressing the + and – buttons and pressing the “Refresh Graph” button. Does it appear to fit these data better? What does that tell you? Does the observed chi-square value appear to be extreme compared to our simulated data?
5.2.5 Chi-Square Test

Finally, click the “Chi-Square Test” button to perform the chi-square test. This will output a graph similar to the graph on the right (Figure 80). What does this tell you about our data? Does there appear to be a significant difference between the groups? Take note that the console contains information about our observed data, our expected data, and the chi-square cell contributions as in Output 18:

```
$`Observed Data`
  Opt-In  Opt-Out  Neutral
Donor  23      41      44
Not    32       9       12

$`Expected Data`
  Opt-In  Opt-Out  Neutral
Donor 36.89441 33.54037 37.56522
Not  18.10559 16.45963 18.43478

$`Chi-Square Cell Contributions`
  Opt-In  Opt-Out  Neutral
Donor  5.232625 1.659076 1.102254
Not 10.662708 3.380759 2.246103
```

Figure 80: Chi-Square Distribution Plot
6. r x k Tables Applet

6.1 Instructor Guide

6.1.1 Applet Introduction

This applet will allow students to explore r by k categorical data through simulation. In order to prepare the applet, students simply double click on the “rbyk.RData” workspace, which launches the R environment on computers with R installed and loads the appropriate functions. To launch the applet, students simply need to launch the function by typing “rbyk()” as in Figure 81. In this instructor guide, as an example, we will be using a fictitious data set with four predictors and three responses.

![Figure 81: Launching the rxk Applet](image)

Once the function is launched, the function will automatically download and load the required packages, as well as load the appropriate data file. Once this completes (which should take no more than a few seconds), the following panel will open as in Figure 82:

![Figure 82: rxk Tables Panel](image)
The applet is designed in such a way to encourage students to help students through their discovery of the data and the relevant statistical concepts. That is, the applet encourages students to explore the data with the Descriptive Statistics button, run simulations of the data to explore associations, and finally, view the theoretical analysis of the data through a Chi-Square test and compare it to the simulated data.

### 6.1.2 Reading In Data

Students must first read in data to R by clicking the “Open Data File” button. This will open the following panel (Figure 83) that will allow students to open either Tab, Comma, or Space delimited files. File type does not matter, as long as the delimiters are consistent. In addition, students can choose data from the clipboard in order to import data from packages like Minitab and JMP. When students click the “Open File” button, they can browse to their data file they wish to analyze. Ensure the data file follows the following requirements:

- First column – response variable
- Second column – explanatory variable
- Third column – count

When the data is read in, the table of the data is shown in the console, as well as the chi-square observed statistic for use in the chi-square simulation as in Output 19:

```
[1] "Your data is:"
   Explanatory
Response Exp1 Exp2 Exp3 Exp4
  Resp1 424 398  39 238
  Resp2  39  94  38 388
  Resp3  48 738 378 292
Chi-Square Test Statistic
  1246.839
```

**Output 19: Original Data**

### 6.1.3 Sum of Pairwise |Differences|

If students click on the “Sum of Pairwise |Differences| button, students will see the following panel appear as in Figure 84:

![Figure 84: Sum of Pairwise |Differences| Panel](image)
If students start the simulation with one repetition, students will be shown the following graph (Figure 85).

![Graph showing original and simulated data with responses by explanatory variables.](image)

**Figure 85: One Repetition**
If students continue to click the shuffle button, they may add repetitions to the simulation and watch the animation of the additional shuffles. Upon adding 9 more repetitions to the simulation, students will see something similar to Figure 86:

![Figure 86: Ten Repetitions](image)

Please note that upon changing values in the R panel (such as the observed statistics), students must press the “Enter” key for the change to be added to R.

From here, students can see a plot of the original data, a plot of the data for a simulation of no association, and a histogram of the simulated pairwise $|differences|$ statistic.

In addition, an animation feature has been implemented so that students may watch what occurs with each shuffle. If students enter a number of repetitions that is 50 or less, the applet will automatically pause for one-half of a second between each shuffle to display the simulated data. If students enter a number of repetitions greater than 50, then the applet will automatically progress to the final simulated data. Now, with 1000 simulations:
Figure 87: rxk with 1000 Simulations
Students can change the observed values of statistics in order to notice how different the observed data is from the simulated data. For example, using the observed value of 1089 and pressing the “Refresh Graph” button yields something similar to Figure 88:

Figure 88: Adding Observed Statistic

From this, students can view how different our observed values of the statistics compare to the rest of the simulated data and get an idea how “extreme” the data are.

Examples of questions to ask students are:

1. Do the stacked bar charts look the same between the original data and the new shuffled data?
2. How do the bar charts change with each shuffle?
3. What do you notice about the histograms of the sum of pairwise differences statistics?
4. How does the histogram look after shuffling 1000 times?
5. Change the observed value of the observed statistic from the Descriptive Statistics and click the “Refresh Graph” button. Where does the observed value lie on the histogram? Is it located more toward the center of the distribution, or is it closer to the tails? How large is the p-value? What does this tell us about the observed value of the statistic?
6.1.4 Sum of $|\text{Phati-Phat}|$

Similar to the previous statistic, if students click on the “Sum of $|\text{Phati-Phat}|$” button, students will see the following panel appear (Figure 89):

![Figure 89: Sum of $|\text{Phati-Phat}|$ Panel](image)

If students start the simulation with one repetition, students will be shown the following graph as in Figure 90:

![Figure 90: One Repetition Graphics Device](image)
If students continue to click the shuffle button, they may add repetitions to the simulation and watch the animation of the additional shuffles. Upon adding 9 more repetitions to the simulation, students will see something similar to Figure 91:

![Figure 91: Ten Repetitions](image)

From here, students can see a plot of the original data, a plot of the data for a simulation of no association, and a histogram of the simulated \(|\text{Phati-Phat}|\) statistics.

In addition, an animation feature has been implemented so that students may watch what occurs with each shuffle. If students enter a number of repetitions that is 50 or less, the applet will automatically pause for one-half of a second between each shuffle to display the simulated data. If students enter a number of repetitions greater than 50, then the applet will automatically progress to the final simulated data. Now, with 1000 simulations:
Figure 92: 1000 Repetitions
Students can change the observed values of statistics in order to notice how different the observed data is from the simulated data. For example, using the observed value of 0.867 and pressing the “Refresh Graph” button yields something similar to Figure 93:

![Figure 93: Changing Observed Value](image)

From this, students can view how different our observed values of the statistics compare to the rest of the simulated data and get an idea how “extreme” the data are. Note that the histogram is quite far from the observed value, which shows just how extreme our data are.

Examples of questions to ask students are:

1. Do the stacked bar charts look the same between the original data and the new shuffled data?
2. How do the bar charts change with each shuffle?
3. What do you notice about the histograms of the sum of pairwise differences statistics?
4. How does the histogram look after shuffling 1000 times?
5. Change the observed value of the observed statistic from the Descriptive Statistics and click the “Refresh Graph” button. Where does the observed value lie on the histogram? Is it located more toward the center of the distribution, or is it closer to the tails? How large is the p-value? What does this tell us about the observed value of the statistic?
6.1.5 Chi-Square Statistic

The chi-square statistic simulation has a few extra features to enable students to explore the distribution as in Figure 94:

The bottom left histogram attempts to plot a normal distribution with mean and standard deviation based on the shuffled data to show students that a normal distribution would not be appropriate for these data. The bottom right graph plots a more appropriate chi-square distribution over the data, while allowing students to explore the distribution through degrees of freedom. That is, students can increase and decrease the degrees of freedom through buttons on the simulation panel as in Figure 95:
If we incorrectly increase the degrees of freedom to 2 and press the Refresh Graph button, we notice that the fit of the curve is much worse than the previous fit as in Figure 96:
However, if students correctly increase the degrees of freedom to 6 and press the Refresh Graph button, students can see that the distribution is much more appropriate for these data (Figure 97):

![Figure 97: Correct Degrees of Freedom](image)

Finally, if students press the Show Last Randomized Table button, students can see what the last shuffled table is that the bar graph is representing and see how it compares to that of the original data (Output 20):

```
<table>
<thead>
<tr>
<th>V1</th>
<th>Exp1</th>
<th>Exp2</th>
<th>Exp3</th>
<th>Exp4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resp1</td>
<td>180</td>
<td>411</td>
<td>172</td>
<td>336</td>
</tr>
<tr>
<td>Resp2</td>
<td>89</td>
<td>221</td>
<td>81</td>
<td>168</td>
</tr>
<tr>
<td>Resp3</td>
<td>242</td>
<td>598</td>
<td>202</td>
<td>414</td>
</tr>
</tbody>
</table>
```

Output 20: Last Shuffled Table Data

In addition, Output 21 shows the p-value to the console based on the chi-square distribution with the degrees of freedom the students have chosen:

```
<table>
<thead>
<tr>
<th>X-squared</th>
<th>df</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1246</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>
```

Output 21: Chi-Square Test - Simulated Data
### 6.1.6 Chi-Square Test

Finally, after exploring different simulations, students can explore the results of a chi-square test. If students click the Chi-Square Test button, students can see a shaded graph of the distribution and p-value similar to Figure 98:

*Figure 98: Chi-Square Distribution Plot*
Relevant information about the data will also be outputted to the R console similarly to Output 22 including the Observed Data, Expected Counts, and Chi-Square Cell Contributions:

```markdown
$`Observed Data`
Explanatory
Response | Exp1   | Exp2   | Exp3   | Exp4   
Resp1    | 424    | 398    | 39     | 238    
Resp2    | 39     | 94     | 38     | 388    
Resp3    | 48     | 738    | 378    | 292    

$`Expected Data`
Explanatory
Response | Exp1   | Exp2   | Exp3   | Exp4   
Resp1    | 180.34329 | 434.0944 | 160.57964 | 323.9827  
Resp2    | 91.73057    | 220.7996  | 81.67791  | 164.7919  
Resp3    | 238.92614    | 575.1060  | 212.74245 | 429.2254  

$`Chi-Square Cell Contributions`
Explanatory
Response | Exp1   | Exp2   | Exp3   | Exp4   
Resp1    | 329.197685 | 3.001206 | 92.051576 | 22.819177 
Resp2    | 30.311739   | 72.817800 | 23.357106 | 302.331912 
Resp3    | 152.569287  | 46.138391 | 128.371448 | 43.871630 
```

Output 22: Chi-Square Test Information

Some suggested questions for students are:

- How does your simulated data change with each “shuffle” of your observed data? How does this compare to your observed data?
- Would a normal approximation be appropriate for these data? Why or why not?
- What happens to the chi-square distribution as you increase the degrees of freedom? As you decrease the degrees of freedom? Which provide a better “fit” to your shuffled data? What value fits your shuffled data best? Why do you think this is?
- What observed value of your statistic would be appropriate to use for these data?
- How would your analysis change if you have twice the number people in each category, but the proportion stays the same? How will your p-value change?
6.2 Student Guide

6.2.1 Applet Introduction

This applet will allow you to explore categorical data through simulation. In order to prepare the applet, double click on the “rbyk.RData” file. To launch the applet, simply launch the function by typing “rbyk()” as in Figure 99. It may take a few seconds to launch, as it may have to download and install a few files before it opens.

When you launch the function, the panel in Figure 100 will open:

Start by clicking “Open Data File” to open the file that contains the data you would like to observe. You may read in data that has Tabs, Commas, or Spaces between the data values (open the data file to tell). You can also open data from the clipboard if you have copied it from programs like Minitab or JMP.

Make sure that your data file has:

- First column – response variable
- Second column – explanatory variable
- Third column – counts

This guide will be using a fictitious data set for demonstration purposes. When you open your data file, the console will display a table of your data, as well as our observed chi-square test statistic to take note of for use in a few minutes as in Output 23:

```
[1] "Your data is:"
Explanatory
Response Exp1 Exp2 Exp3 Exp4
Resp1 424 398 39 238
Resp2 39 94 38 388
Resp3 48 738 378 292
Chi-Square Test Statistic
  1246.839
```

Output 23: Original Data Information

6.2.2 Sum of Pairwise |Differences|

Start by pressing the Sum of Pairwise |Differences| button to open the panel in Figure 101. From here, press the “Shuffle” button to start shuffling your data.
Once you shuffle the data for the first time, take note of the simulated version of the data. How do those barcharts compare to the barcharts in our original data? Does the relationship between the two groups change?

Continue by changing the number of shuffles to 10. (You have to press the “Enter” key when changing any values to have the applet accept the change.) From the animation, how does the relationship between the two groups change? Is there a large change in the barcharts with each shuffle?

Add another 40 shuffles to get to 50 total. Take note of the histogram of your observed statistic. Is the histogram symmetric? What value appears the most? Do these make sense with our data?

Add another 950 shuffles and take a look at the final histogram in Figure 102. Try changing the observed value of that statistic to the value for the original data and click the “Refresh Graph” button. How extreme is that value? What does that tell you about the association in the context of our data?

You may also click the checkbox “Show the Last Randomized Table,” which will show the last simulated table to see the upper right bar chart expressed as a table. How does it compare to that of the original data?

### 6.2.3 Sum of |Phati-Phat|

If you click on the “Sum of |Phati-Phat|” button, you will see a similar panel to the previous simulation panel. From this panel, you can run a similar test to the previous one. Go ahead and repeat the same procedure from above. How does this statistic differ? Does it provide similar results to the ones you obtained above?

### 6.2.4 Chi-Square Simulation

Now moving on to the “Chi-Square Simulation” panel, you should see the panel in Figure 103 appear. Feel free to follow the same procedure as before to show a graph similar to Figure 104.
Does the normal distribution in the lower left corner appear appropriate for these data? Once you run 1000 repetitions, try adjusting the degrees of freedom of the green chi-square distribution curve in the lower right panel by pressing the + and – buttons and pressing the “Refresh Graph” button. Does it appear to fit these data better? What does that tell you? Does the observed chi-square value (from the console when you opened your data) appear to be extreme compared to our simulated data?

### 6.2.5 Chi-Square Test

Finally, click the “Chi-Square Test” button to perform the chi-square test. This will output a graph similar to the graph below in Figure 105. What does this tell you about our data? Does there appear to be a significant difference between the groups? Take note that the console contains information about our observed data, our expected data, and the chi-square cell contributions as in Output 24:

```r
$`Observed Data'
   Explanatory
   Response Exp1 Exp2 Exp3 Exp4
   Resp1 424 398 39 238
   Resp2 39 94 38 388
   Resp3 48 738 378 292

$`Expected Data'
   Explanatory
   Response Exp1 Exp2 Exp3 Exp4
   Resp1 180.34329 434.0944 160.57964 323.9827
   Resp2 91.73057 220.7996 81.67791 164.7919
   Resp3 238.92614 575.1060 212.74245 429.2254

$`Chi-Square Cell Contributions'
   Explanatory
   Response Exp1 Exp2 Exp3 Exp4
   Resp1 329.197685 3.001206 92.051576 22.819177
   Resp2 30.311739 72.817800 23.357106 302.331912
   Resp3 152.569287 46.138391 128.371448 43.871630
```

Figure 105: Chi-Square Distribution Plot
7. Analysis of Variance Applet

7.1 Instructor Guide

7.1.1 Applet Introduction

This applet will allow students to explore simple linear regression through simulation. In order to prepare the applet, students simply double click on the “One Way ANOVA.RData” workspace, which launches the R environment on computers with R installed and loads the appropriate functions. To launch the applet, students simply need to launch the function by typing “onewayanova()” as in Figure 106. In this instructor guide, as an example, we will be using diet type (Atkins, Weight Watchers, Zone, and Ornish) to predict weight loss.

Once the function is launched, the function will automatically download and load the required packages, as well as load the appropriate data file. Once this completes (which should take no more than a few seconds), the panel in Figure 107 will open:
The applet is designed in such a way to encourage students to help students through their discovery of the data and the relevant statistical concepts. That is, the applet encourages students to explore the data with the Descriptive Statistics button, run simulations of the data to explore associations, and finally, view the theoretical analysis of the data through ANOVA and compare it to the simulated data.

7.1.2 Reading In Data
Students must first read in data to R by clicking the “Open Data File” button. This will open the panel in Figure 108 that will allow students to open either Tab, Comma, or Space delimited files. File type does not matter, as long as the delimiters are consistent. When students click the “Open File” button, they can browse to their data file they wish to analyze. Ensure the data file follows the following requirements:

- First column - observation number
- Second column – group
- Third column – quantitative response

7.1.3 Descriptive Statistics
Beginning with the descriptive statistics function, students will see a graph similar to Figure 109:
In addition, R will output the following information to the console (Output 25):

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atkins</td>
<td>21</td>
<td>3.92</td>
<td>6.05</td>
</tr>
<tr>
<td>Zone</td>
<td>20</td>
<td>6.56</td>
<td>9.29</td>
</tr>
<tr>
<td>Ornish</td>
<td>26</td>
<td>4.59</td>
<td>5.39</td>
</tr>
<tr>
<td>WeightWatchers</td>
<td>26</td>
<td>4.88</td>
<td>6.92</td>
</tr>
<tr>
<td>Total</td>
<td>93</td>
<td>4.95</td>
<td>6.89</td>
</tr>
</tbody>
</table>

$`\text{Observed Value of } \sum(y_{ibar}-y_{bar})^2`$

[1] 3.788785

$`\text{Observed Value of } F`$

[1] 0.5361475

**Output 25: Original Data Information**

In this way, students may begin exploring their visual understanding of the data through the graph and plotting of the regression line. In addition, they can obtain the observed values of $\sum(\bar{y}_i - \bar{y})^2$ and F statistics that will be useful for the simulation panel.

At this point, examples of questions to ask students are:

- What do you observe?
- Do the boxplots look the same between all of the groups?
- Are there any unusual observations?
- How would you characterize the amount of overlap between the distributions?

In addition, it would be beneficial to look at other samples with more and less variability and show how they would compare the strength of evidence in each case.
7.1.4 Simulate No Association

If students click on the Simulate No Association button, students will see the panel from Figure 110 appear:

![Simulate No Association Panel](image)

Figure 110: Simulate No Association Panel

In addition, a graphics device similar to Figure 111 will open:

![One Repetition](image)

Figure 111: One Repetition
If students continue to click the shuffle button, they may add repetitions to the simulation and watch the animation of the additional shuffles. Upon adding 9 more repetitions to the simulation, students will see something similar to Figure 112:

![Figure 112: 10 Repetitions](image)

Please note that upon changing values in the R panel (such as the observed statistics), students must press the “Enter” key for the change to be added to R.

From here, students can see a plot of the original data, a plot of the data for a simulation of no association, a histogram of the simulated $\sum (y_i - \bar{y})^2$ statistic, and a histogram of the simulated F statistic under $H_0$. 


In addition, an animation feature has been implemented so that students may watch what occurs with each shuffle. If students enter a number of repetitions that is 50 or less, the applet will automatically pause for one-half of a second between each shuffle to display the simulated data. If students enter a number of repetitions greater than 50, then the applet will automatically progress to the final simulated data. For example, with 1000 simulations, students will see a graph similar to Figure 113:

![Graphs showing simulation results with 1000 repetitions](image-url)

Figure 113: 1000 Simulations
Students can change the observed values of statistics in order to notice how different the observed data is from the simulated data. For example, using the observed values of: \( \sum (\bar{y}_i - \bar{y})^2 = 3.789 \) and \( F = 0.536 \) and pressing the “Refresh Graph” button results in a graph similar to Figure 114:

![Figure 114: Changing Observed Values](image)

From this, students can view how different our observed values of the statistics compare to the rest of the simulated data and get an idea how “extreme” the data are. In addition, students can add and subtract degrees of freedom for the F distribution. After adjusting them to the appropriate degrees of freedom (3 numerator df and 89 denominator df), the curve changes to a much more appropriate fit as in Figure 115:
Examples of questions to ask students are:

6. Do the boxplots look the same between the original data and the new shuffled data?
7. How do the box plots change with each shuffle?
8. What do you notice about the histograms of the $\sum (\bar{y}_i - \bar{y})^2$ and F statistics? Are they symmetric?
9. How do the histograms look after shuffling 1000 times?
10. Try adding and subtracting numerator and denominator degrees of freedom to make the yellow curve fit our simulated statistics better. What happens as you increase and decrease the degrees of freedom? Make sure to change the degrees of freedom values with the buttons and click the “Refresh Graph” button to update the curve.
11. Change the observed value of the $\sum (\bar{y}_i - \bar{y})^2$ and F statistics from the Descriptive Statistics and click the “Refresh Graph” button. Where do the observed values lie on the histograms? Is it located more toward the center of the distribution, or is it closer to the tails? How large are the p-values? What does this tell us about the observed value of the statistics?
7.1.5 ANOVA / F Test

Finally, if students press the “ANOVA / F Test” button, students can explore the assumptions of the ANOVA model, as well as the ANOVA table and assumptions output. In our example, the following model data is outputted to the console:

ANOVA Table:

Analysis of Variance Table

Response: diet

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>weightloss</td>
<td>3</td>
<td>77.6</td>
<td>25.866</td>
<td>0.5361</td>
<td>0.6587</td>
</tr>
<tr>
<td>Residuals</td>
<td>89</td>
<td>4293.7</td>
<td>48.244</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Output 26: Analysis of Variance Information

Shapiro-Wilk Test of Normality:

Shapiro-Wilk normality test

data: lm.obj$residuals
W = 0.9627, p-value = 0.009277

Output 27: Testing Normality of the Residuals

Bartlett Test of Homogeneity of Variances:

Bartlett test of homogeneity of variances

data: dataset
Bartlett's K-squared = 209.9793, df = 1, p-value < 2.2e-16

Output 28: Testing Homoscedasticity
Similarly, students can view the graphs found in Figure 116:

Here, students can see the boxplots of the original data by diet group, a plot of the standardized residuals vs. fits in order to determine violations of homoscedasticity and view any outliers, a normal quantile-quantile plot to determine violations of normality of the residuals, and the F distribution output with p-value and critical value.

Here, examples of questions for students are:

1. Where is the F distribution critical value of the data located on the curve? How large does that make the p-value? What does this tell you about these populations?
2. Based on the Residuals vs. Fits plot and the Normal Q-Q plot, does it look like the normality or constant variance assumptions are violated? Check your answer with the p-values to the “Shapiro-Wilk normality test” and the “Bartlett test of homogeneity of variances” tests in the console. What do these tests tell you?
7.2 Student Guide

7.2.1 Applet Introduction

This applet will allow you to visualize data with categorical groups and quantitative responses through Analysis of Variance (ANOVA). In order to prepare the applet, double click on the “One Way ANOVA.RData” file, which launches the program. To launch the applet, simply launch the function by typing “onewayanova()” into the console as in Figure 117. It may take a few seconds to launch, as it may have to download and install a few files before it opens.

When you launch the function, the following panel will open (Figure 118):

Start by clicking “Open Data File” to open the file that contains the data you would like to observe.

Make sure that your data file has:

- First column - observation number
- Second column – group
- Third column – quantitative response
- No missing observations.

Then, if you click the “Descriptive Statistics” button, you will see a graph of your data. What do you observe? Do the boxplots look the same between all of the groups? Are there any unusual observations? Take note of the two observed statistics in the console. These will be important in just a moment.
7.2.2 Simulate No Association

If you click the “Simulate No Association” button, you will see the following panel open (Figure 119):

Start by shuffling 1 repetition.

Do the boxplots look the same between the original data and the new shuffled data? Add 9 more shuffles to make 10 total. (Note that whenever you change a value, you must hit “Enter” to save the new value.) Make sure to watch the animation of the graph of the shuffled data. How do the box plots change with each repetition?

Now add 40 more shuffles to make 50 total. What do you notice about the histograms? Are they symmetric? Now add 950 more shuffles to make 1000 total. How do the histograms look now?

Change the observed value of the $\sum(y_i - \bar{y})^2$ and F statistics from the Descriptive Statistics and click the “Refresh Graph” button. Where do the observed values lie on the histograms? Is it located more toward zero, or is it closer to the right tail? How large are the p-values? What does this tell us about the observed value of the statistics?

7.2.3 ANOVA / F Test

Comparing the theoretical answer to our simulated answer for the data, to look at the ANOVA output, click the “ANOVA / F Test” button. This will open several graphs and display output to the console. Check out the graph of the F-curve. Where is the critical value of the data located? How large does that make the p-value? What does this tell you about these groups?

Take note that the histogram for the F statistic you see here can be approximated by an F curve with the listed degrees of freedom on the previous Simulate No Association panel.

Based on the Residuals vs. Fits plot and the Normal Q-Q plot, does it look like the normality or constant variance assumptions are violated? Check your answer with the p-values to the “Shapiro-Wilk normality test” and the “Bartlett test of homogeneity of variances” tests in the console.
8. Simple Linear Regression Applet

8.1 Instructor Guide

8.1.1 Applet Introduction

This applet will allow students to explore simple linear regression through simulation. In order to prepare the applet, students simply double click on the Correlation.RData workspace, which launches the R environment on computers with R installed and loads the appropriate functions. To launch the applet, students simply need to launch the function by typing “correlation()” as in Figure 120. In this instructor guide, we will be using the number of bedrooms to predict rent prices.

![Figure 120: Launching the Regression Applet](image)

Once the function is launched, the function will automatically download and load the required packages, as well as load the appropriate data file. Once this completes (which should take no more than a few seconds), the following panel will open as in Figure 121:

![Figure 121: Correlation Applet](image)
The applet is designed in such a way to encourage students to help students through their discovery of the data and the relevant statistical concepts. That is, the applet encourages students to explore the data with the Descriptive Statistics button, run simulations of the data to explore associations, and finally, view the full simple linear regression of the data.

8.1.2 Reading In Data

Students must first read in data to R by clicking the “Open Data File” button. This will open the following panel (Figure 122) that will allow students to open either Tab, Comma, or Space delimited files. When students click the “Open File” button, they can browse to their data file they wish to analyze. Ensure the data file follows the following requirements:

- First column – response variable
- Second column – explanatory variable

8.1.3 Descriptive Statistics

Beginning with the descriptive statistics function, students will see the following graph:

![Original Data Graph](image)

Figure 123: Regression Graph
In addition, R will output the following information to the console as in Output 29:

```
$'Correlation Coefficient r'
[1] 0.8360753

$'Simple Linear Regression'

Call:
  lm(formula = dataset)

Coefficients:
  (Intercept)     Bedrooms
  353.8        402.8

$'Observed Value of Slope, b1'
  Bedrooms
  402.7847
```

Output 29: Original Data Regression Information

In this way, students may begin exploring their visual understanding of the data through the graph and plotting of the regression line. In addition, they can interpret the correlation coefficient of the data, and obtain the regression formula that will be useful for the simulation panel.

8.1.4 Simulate No Association

If students click on the Simulate No Association button, students will see the following panel appear as in Figure 124:
If students start the simulation with one repetition, students will be shown a graph similar to Figure 125.

Figure 125: One Repetition
If students continue to click the shuffle button, they may add repetitions to the simulation. Upon adding 9 more repetitions to the simulation, students will see something similar to Figure 126:

![Figure 126: 10 Repetitions](image)

Please note that upon changing values, students must press the “Enter” key for the change to be added to R.

From here, students can see a plot of the original data with regression line, a plot of the data for a simulation of no association, a histogram of the simulated correlation coefficient, and a histogram of the simulated slope.

In addition, an animation feature has been implemented so that students may watch what occurs with each shuffle. If students enter a number of repetitions that is 50 or less, the applet will automatically pause for one-half of a second between each shuffle to display the simulated data. If students enter a number of repetitions greater than 50, then the applet will automatically progress to the final simulated data. For example, with 1000 simulations, students will see a graph similar to Figure 127:
Figure 127: 1000 Repetitions
Students can change the observed values of statistics in order to notice how different the observed data is from the simulated data. For example, using the observed values of: \( r = 0.836 \) and \( b_1 = 402.8 \) and pressing the “Refresh Graph” button yields a graph similar to Figure 128:

**Figure 128: Changing Observed Statistics**
8.1.5 Simple Linear Regression

Finally, if students press the “Simple Linear Regression” button, students can explore the assumptions of linear regression, as well as the model output. In our example, the following model data is outputted to the console:

Regression Model Information:

```
Call:
  lm(formula = dataset)

Residuals:
   Min      1Q  Median      3Q     Max
-614.99 -164.29  -35.81  239.89 1135.01

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
(Intercept)   353.85     106.76   3.314  0.00156 **
   Bedrooms     402.78      34.12  11.805  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 324.1 on 60 degrees of freedom
Multiple R-squared: 0.699,    Adjusted R-squared: 0.694
F-statistic: 139.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Output 30: Simple Linear Regression

ANOVA Table:

```
Analysis of Variance Table

Response: Rent

   Df Sum Sq Mean Sq  F value     Pr(>F)
Bedrooms 1 14635216 14635216 139.35 < 2.2e-16 ***
Residuals 60  6301489   105025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Output 31: Analysis of Variance Information

Shapiro-Wilk Test of Normality:

```
Shapiro-Wilk normality test

data:  slr$residuals
W = 0.9675, p-value = 0.09917
```

Output 32: Testing Normality of Residuals
Similarly, students can view the following graphs in Figure 129:

![Figure 129: Regression Diagnostics](image)

Here, students can see the regression line fitted with the original data, a plot of the residuals vs. fits in order to determine violations of homoscedasticity and view any outliers, a normal quantile-quantile plot to determine violations of normality of the residuals, and a residual vs. order plot to determine violations of independence.
8.2 Student Guide

8.2.1 Applet Introduction

This applet will allow you to visualize data with quantitative explanatory and response through Simple Linear Regression (SLR). In order to prepare the applet, double click on the “Correlation.RData” file, which launches the program. To launch the applet, simply launch the function by typing “correlation()” into the console as in Figure 130. It may take a few seconds to launch, as it may have to download and install a few files before it opens.

When you launch the function, the following panel will open (Figure 131):

Start by clicking “Open Data File” to open the file that contains the data you would like to observe.

Make sure that your data file has:

- First column – quantitative response
- Second column – quantitative predictor
- No missing observations.

Then, if you click the “Descriptive Statistics” button, you will see a graph of your data with the regression line in red. What do you observe? Does there appear to be an association? If so, what direction does the association appear to be following? How much variability around the regression line do the points seem to have? Does there appear to be any unusual observations? Take note of the two observed statistics in the console. These will be important in just a moment.
8.2.2 Simulate No Association

If you click the “Simulate No Association” button, you will see the following panel open (Figure 132):

Start by shuffling 1 repetition.

Do the graphs and regression lines look the same between the original data and the new shuffled data? How has the association changed in the new shuffled data? Add 9 more shuffles to make 10 total. (Note that whenever you change a value, you must hit “Enter” to save the new value.) Make sure to watch the animation of the graph of the shuffled data. How does the regression change with each shuffle?

Now add 40 more shuffles to make 50 total. What do you notice about the histograms? Are they symmetric? Now add 950 more shuffles to make 1000 total. How do the histograms look now? What value are they centered around? Does this make sense?

Change the observed value of the correlation coefficient $r$ and slope $b_1$ from the Descriptive Statistics output and click the “Refresh Graph” button. Where do the observed values lie on the histograms? Is it located more toward the center of the distribution, or is it closer to the tails? Why are there two p-values? How large are the p-values? What does this tell us about the observed value of the statistics?

8.2.3 Simple Linear Regression

Comparing the theoretical answer to our simulated answer for the data, to look at the Simple Linear Regression output, click the “Simple Linear Regression” button. This will open several graphs and display output to the console. Check out the F-statistic and the p-value in the console. Is this significant? What does this tell you about this association in the context of our data?

Based on the Normal Q-Q plot, does it look like the normality assumption is violated? Check your answer with the p-value to the “Shapiro-Wilk normality test” in the console. Looking at the Residuals vs. Fits plot, does the constant variance (homoscedasticity) assumption appear to be satisfied? How do you know? Finally, looking at the Residual vs. Order plot, are there any issues with independence?
9. Feedback from Students

Several of these applets have since been used in introductory statistics courses. Students used the 2x3 and ANOVA applets in Drs. Chance and McGaughey’s STAT 217 classes in Winter and Spring 2012. Also, other students used the ANOVA applet in Dr. Roy’s STAT 302 Lecture and Investigation assignment.

During these labs, after overcoming the initial hurdle of opening the applet, students all appreciated the visual nature of the applet. In talking with students, most felt they understood the concepts presented much better due to the applet.

In addition, a survey was distributed to Dr. Roy’s STAT 302 classes following use of the ANOVA applet during lecture and an investigation assignment. The survey posed the following six questions to the students:

1. How did you feel about working with the R package compared to other software like Minitab?
2. Was the applet easy to follow?
3. Do you feel like you understand what the applet is doing? If not, what can be changed to further your understanding?
4. Do you feel that the applet was useful in helping you understand the randomization-based method ANOVA? Why or why not?
5. What else, if anything, would you like the applet to do to help you in your learning of statistics?
6. Would you recommend that statistics professors use the applet in future courses to teach ANOVA?

In particular, this survey showed that students felt the applet was a resounding success. All students recommended that statistics professors use the applets in future statistics courses. In addition, nearly all students felt that the applet was helpful in furthering their knowledge of ANOVA. Several students put the applet on the same level as Minitab in terms of ease of use, and one student even liked the applet better than Minitab for understanding statistics concepts.
10. **Next Steps**

There are several directions that I would like to take the project in the future. First, I would like to build an applet to explore multiple regression concepts, as this can be a challenging concept for students to understand, and especially to imagine in three dimensions. I believe that having an applet would allow students to better visualize what multiple regression is doing. See Figure 133 for an example of a graph that the applet would show students. Also, it would be helpful if I extend the one-way ANOVA applet to a multifactor ANOVA applet. In addition, I believe it would be beneficial if students had the ability to input their own test statistics into the applets to simulate and analyze so that the students are not bound to just the statistics built into the applets.

![Figure 133: Multiple Regression Example Graph](image)

I would also like to build in functionality to embed R in an HTML format so that applets can be run outside of the R interface, and especially on the Internet. This would dramatically open up the possibilities of these applets use in the classrooms and in students’ homes. Next, I would like to build in the functionality to open data from other statistics packages like Minitab and JMP without the need to copy it to the clipboard, as that adds an unnecessary step for students and instructors.

Finally, I would like to build similar functionality in R as found in the Minitab View Probability function, as this is quite a powerful tool for students to use in learning about distributions. However, no similar functionality can be found in R.