TECHNOLOGIES TO FACILITATE EACH STAGE OF STUDENT-DIRECTED STATISTICS PROJECTS

Dianna J. Spence
University of North Georgia
Dahlonega, GA 30597
djspence@ung.edu

Brad Bailey
University of North Georgia
Dahlonega, GA 30597
bbailey@ung.edu

Introduction

Researchers and educators often recommend the use of authentic projects with real data as an effective technique for teaching statistics. We have previously reported on our work with such projects, describing the projects themselves in detail (Bailey, Spence, & Sinn, 2013); suggesting primary technologies for implementing such projects (Spence & Bailey, 2013); and sharing preliminary research findings on the impact of such projects (Spence, Sharp, & Sinn, 2011; Spence & Bailey, 2015). In the present paper, we briefly summarize the nature of the projects, the rationale for implementing them, and the technologies previously emphasized to support them. We then describe additional technologies available to enhance students' performance and/or understanding of various project tasks, noting the potential added benefits of these technologies.

Summary of Projects and Prior Reports

Consistent with recommendations that statistics education be student-centered (Roseth, Garfield, & Ben-Zvi, 2008) and foster authentic experience with statistical inquiry (Bryce, 2005), the projects we describe here are highly student-directed: Students select variables, research question, and sampling strategies; carry out their own collection, organization, and analysis of the data; and report their research design, methods, and findings in a written report and in a formal presentation to instructor and peers. Earlier papers have given extensive descriptions of each of these project phases (see Bailey, Spence, & Sinn, 2013). In previous reports, the focus of these projects was limited primarily to linear regression and simple t-test designs. As our work has progressed, the same model has been applied to projects using other types of statistical analysis, including chi-square, z-test for proportions, and ANOVA.

To provide some context, we provide a brief summary here of basic technologies that have previously been suggested to facilitate the project phases of data collection, analysis, and dissemination; see Spence & Bailey (2013) for a more complete discussion of these technologies. First, three broad categories of technology were recommended to support data collection. One such category is comprised of online survey tools, such as

Survey Monkey or Google Docs tools; another category includes specialized technology devices for measuring physical phenomena, such as Texas Instruments CBR™ systems; and the third category encompasses myriad web-based data repositories, with data made available by government agencies, consumer groups, market research firms, sports organizations and franchises, and various industries. Next, to support data analysis, the primary emphasis has been on the TI-83/84 family of calculators and on Microsoft Excel, both of which offer relatively basic functionality to compute standard descriptive statistics, construct simple charts, and conduct common statistical tests. Finally, our previous discussion of technology to support project dissemination has focused primarily on tools for creating the in-class presentation, such as PowerPoint, Adobe, or Prezi.

Additional Technologies and Their Benefits

Data Analysis: More Powerful Tools

One advantage of the calculator and spreadsheet tools emphasized previously is that students often have some exposure to these tools even before using them in a statistics class; they are accessible and familiar, and their use is widespread. However, many software packages designed specifically for statistical analysis offer more functionality, options, and features than do calculators and spreadsheets. Two such software options that we have focused on are JMP and R with RStudio.

JMP is a powerful statistical software package by SAS Corporation; it facilitates data visualization and graphical representations through a very user-friendly Graphical User Interface (GUI) with extremely intuitive menus (see Jones & Sall, 2011). The ease-of-use and emphasis on data visualization make JMP an ideal tool for students learning elementary statistics. Descriptive statistics and accompanying charts are easy to create and rich with representations and connections, as illustrated in Figures 1 and 2.

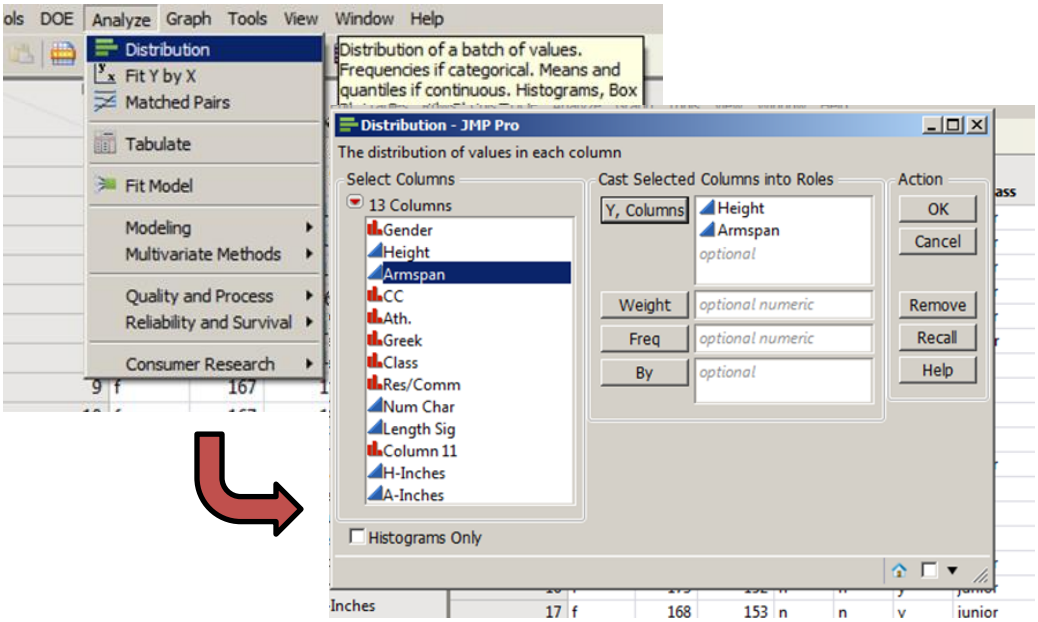

Figure 1. Menu-Driven Creation of Descriptive Statistics and Charts in JMP

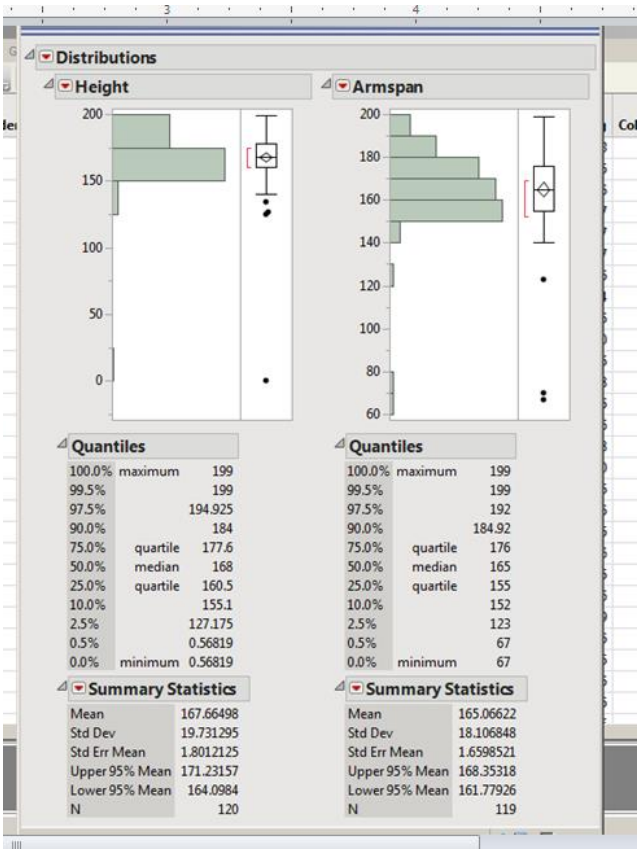

Figure 2. Descriptive Statistics and Data Representation in JMP

JMP is often described as "exploratory" or "discovery" software (e.g., Jones & Sall, 2011). Thus, it stands to reason that the software fosters the exploration of potential relationships among the data as one step in the process toward conducting a statistical test of significance, as illustrated in Figures 3 – 6.

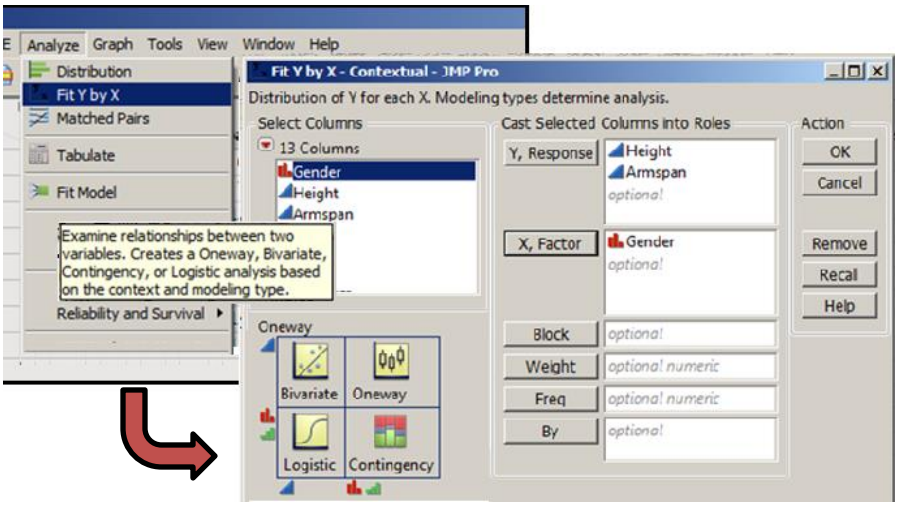

Figure 3. JMP Interface for Exploring Relationships Among Variables

Figures 3 – 4 illustrate the process of comparing the distribution of a quantitative variable in two different groups as one step in the analysis, eventually leading to a t-test for the significance of the observed difference.

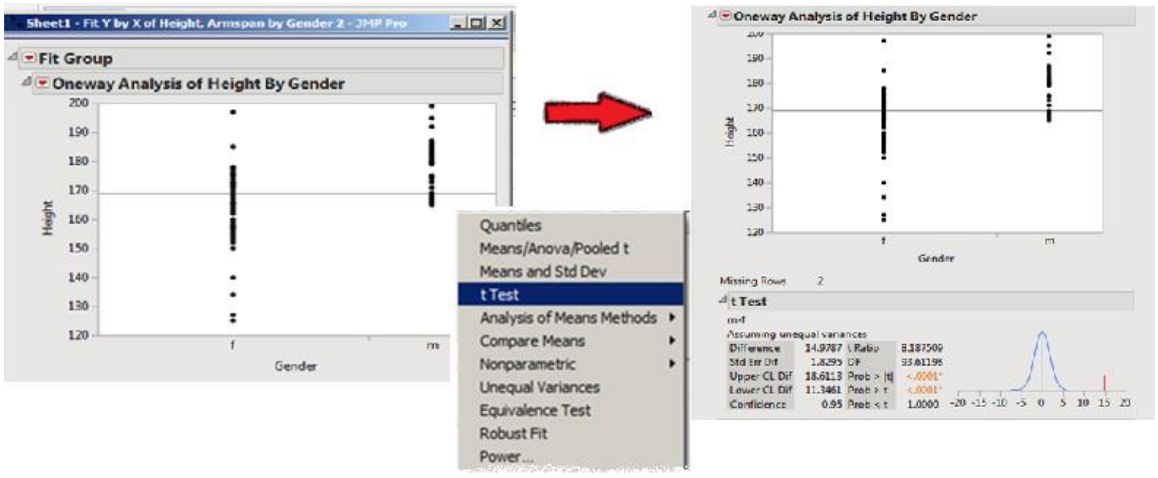

Figure 4. Quantitative Comparison of Two Groups in JMP

Figures 5 – 6 show a similar process with the investigation of two quantitative variables and their relationship to each other, first through graphic representation of the relationship, and then through the appropriate analysis and significance test.

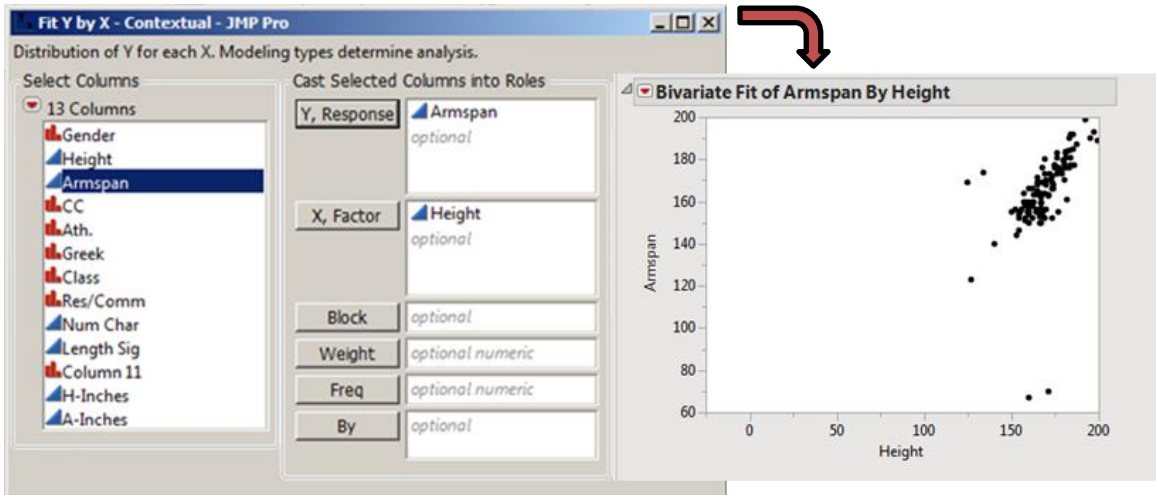

Figure 5. Relationship Between Two Quantitative Variables in JMP

It is interesting to note that far more than calculators or spreadsheets, this software helps to guide the student toward the correct analyses and statistical tests by only offering options that make sense for the variables that are selected. For instructors who wish to focus more on conceptual exploration and the "big picture" of statistical significance, this approach can be very beneficial.

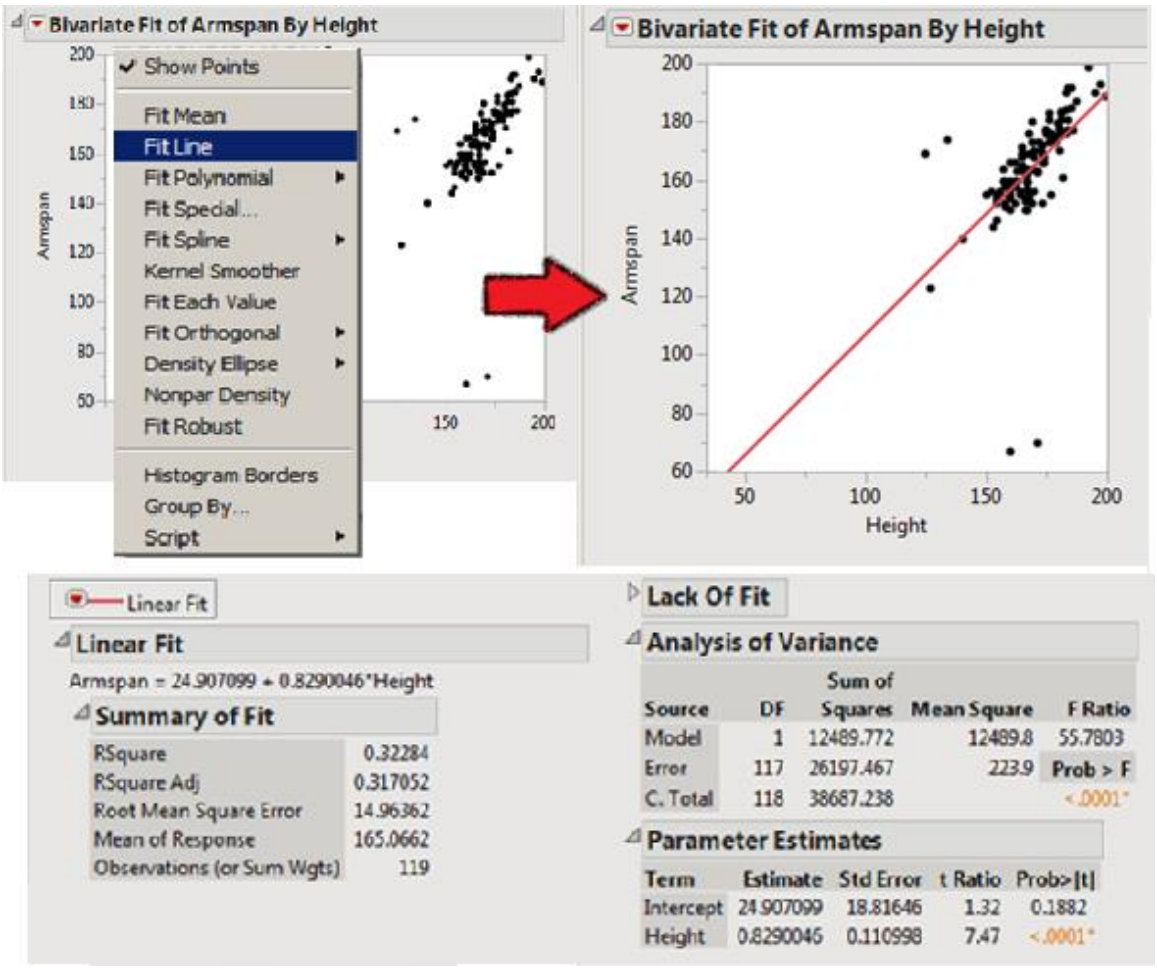

Figure 6. Analysis and Significance Test for Linear Relationship in JMP

Another software option is R, which is a powerful environment for statistical analysis. At its core, R is a command line interface and a specialized programming language, in which students can execute individual instructions or develop scripts at many levels of sophistication. A noteworthy advantage of R is that it is freely available for download on the Internet. The environment can also be extended with packages—libraries of additional functions developed and made available by members of the R user community. These packages are also freely available. One such package designed with students in mind is the Mosaic package, which offers a manageable set of functions which are syntactically consistent and reasonably simple to learn (Pruim, Horton, & Kaplan, 2014). Further, as a helpful alternative to the purely command line interface that is native to R, students can use RStudio, an Integrated Development Environment (IDE) that manages the non-programming aspects of R (e.g., file management). Like the R environment and packages, RStudio is also freely available for download. The IDE does not provide a GUI for the statistical instructions themselves; these must still be typed and adhere to proper syntax. Nevertheless, the RStudio interface in conjunction with packages like Mosaic can make students' project tasks more accessible and vastly reduce their learning curve.

An example of data exploration using R within the RStudio interface is shown in Figure 7. The boxplots compare a quantitative variable (time spent exercising) between military cadets and civilians for each gender. The example illustrates the power and flexibility of a single R command, as well as the organization of the RStudio environment.

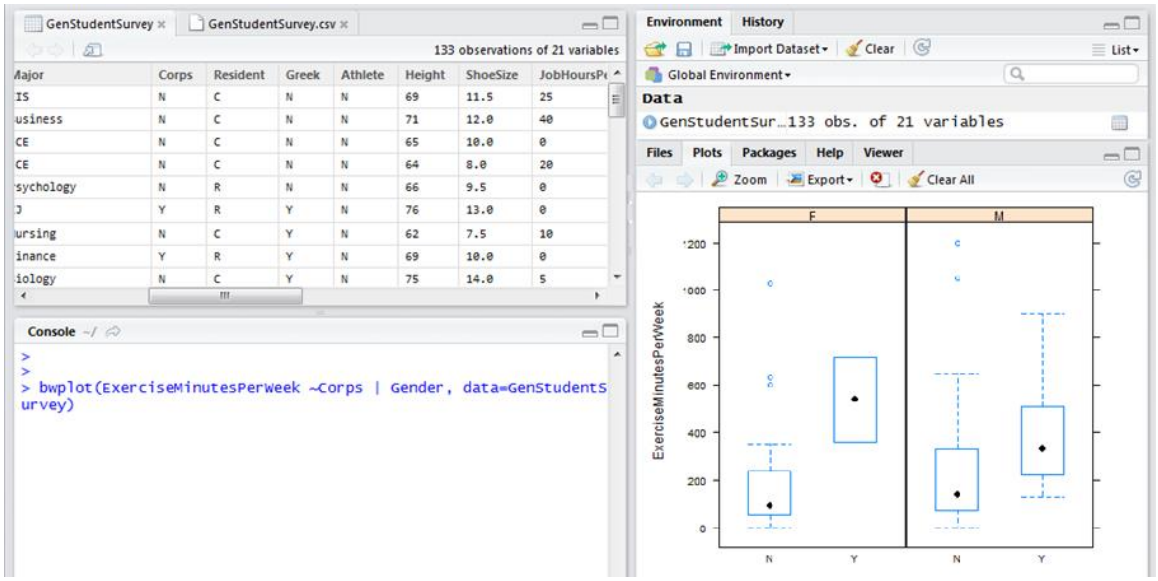

Figure 7. Data Exploration with R/RStudio

When using environments like JMP and R/RStudio to carry out projects, instructors can orchestrate a great deal of exploration and discovery about the data set that each student has collected. Although the JMP interface is intuitive and user-friendly, it will still entail a learning curve and a modest investment of time to get acquainted with it. Many instructors find these a small price to pay for the rich environment, data visualization, and exploration opportunities that result. Likewise, although the RStudio IDE and packages available for R make it a more viable environment for students, they will still require some time to learn their way around. Nevertheless, the freely available environment has much to offer, including R Markdown, an elegant mechanism for formal reporting, as discussed later under *Dissemination*.

Data Analysis: Simulation-Based Inference

Another approach that has recently gained momentum among statistics educators is that of simulation-based inference (SBI). Instead of making a theoretical distribution the primary context for statistical inference, students simulate repeated random samples under assumptions consistent with the null hypothesis to estimate the likelihood of obtaining a sample as extreme as the one observed if the null hypothesis is in fact true. The approach is exceptionally valuable, not only as a conceptual tool for students to grasp the underpinning of inferential reasoning, but also as a viable method of testing for statistical significance (Cobb, 2007; Rossman & Chance, 2014).

We have focused on two tools that facilitate SBI. These are StatKey and the Rossman-Chance applet collection. While both were developed to accompany a specific textbook, both are valuable tools even if used independently of their partner texts. Figure 8 shows an example of one of the Rossman-Chance applets to determine if a difference in means between two groups is significant by shuffling the data points and randomly reassigning each point to one of the two groups—thereby simulating how the samples might have been distributed if the groups were truly no different. From these simulated repeated samples, a distribution of sample statistics (mean differences) can be constructed and used as the basis for inference, as illustrated on the right.

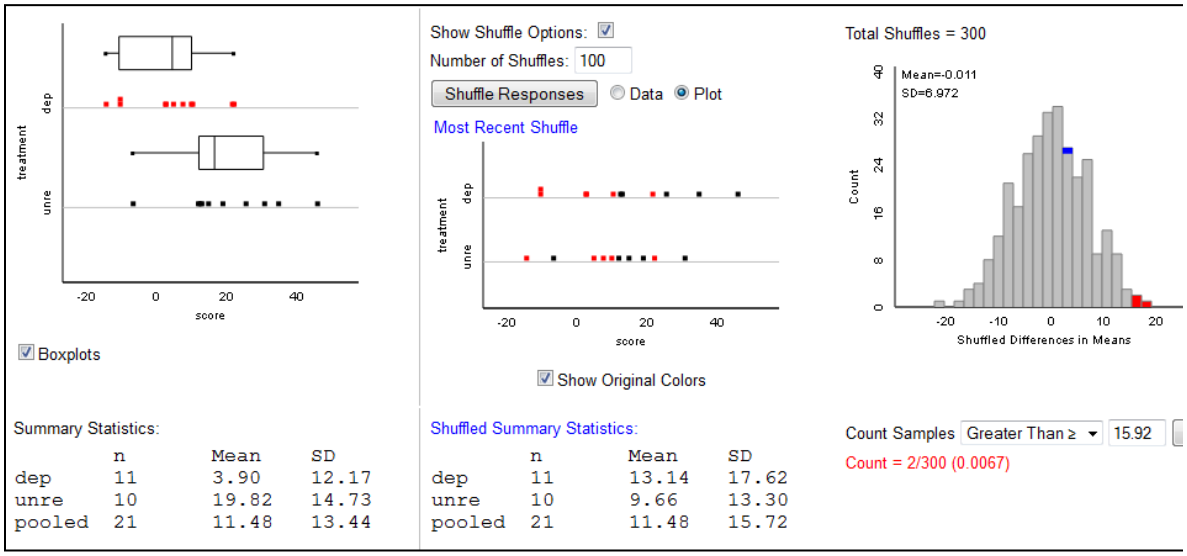

Figure 8. Simulation-Based Inference with one of the Rossman-Chance Applets

It is important to recognize that the student must still use the software to generate and interpret a p-value, but this p-value is based on the simulated samples that could have happened *with the given data set*, rather than on some theoretical distribution, which has numerous assumptions that the data set is required (and often assumed) to meet. Thus, the probabilistic question that the p-value was intended to address is answered most satisfactorily by the p-value generated with this technique.

Many options exist for using one of these SBI tools during the data analysis phase of the project. Depending on departmental or institutional requirements, instructors can have students use SBI instead of the traditional theoretical distribution (in the example above, this would be a t-test.) Alternatively, students could conduct two significance tests— one with SBI and another using the theoretical distribution— and compare their results. In either case, SBI can add a valuable component to the data analysis phase of the project.

Dissemination

Students must ultimately disseminate their work through both a written report and a formal presentation. The most commonly recommended tool for the report has simply been a word processor, and the typical suggestions for the presentation have been

PowerPoint, Adobe, and Prezi. Yet both JMP and the R/RStudio environment have useful functionality for creating reports and presentations. JMP has an Interactive HTML feature that allows students to save their results into an interactive HTML page, which can then be viewed in any browser, without requiring JMP to be present. The interactive component allows the viewer to engage dynamically with the output. For example, the page could allow the viewer to use a slider to see shifts in the data over time or with the change in a particular independent variable. Another possibility would be to allow the viewer to select a subset of the data set to see which portion of a plot corresponds to the designated subset. These features leverage JMP's data visualization capabilities nicely, and the HTML pages created can be used dynamically in a student's presentation with very effective results.

For creating project reports and presentations, RStudio allows students to place their written paragraphs, R commands, statistical output, and tables and charts into a single document using R Markdown, an authoring format for embedding R content into text, with familiar formatting options such as different sized headings, bulleted lists, etc. The result can then be saved in multiple formats, including HTML and PDF. A particular advantage to this means of reporting is that students can easily show exactly how they obtained their results, as the R code itself can be made part of the document.

Conclusion

Many technology options are available to support learning through student-directed statistics projects. Here we have highlighted some additional tools and approaches to supplement those discussed in previous work (Spence & Bailey, 2013). Even when these projects have been implemented with the most basic tools (e.g., Excel), our research suggests that such projects have a positive impact on student outcomes (Spence & Bailey, 2015). The technologies discussed in this paper have the potential to add even more value to these projects.

Acknowledgements

This work is supported by NSF grant award DUE-1021584. Any findings or recommendations presented in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bailey, B., Spence, D. J., & Sinn, R. (2013). Implementation of discovery projects in statistics. *Journal of Statistics Education [online]*, *21*(3). URL: http://www.amstat.org/publications/jse/v21n3/bailey.pdf

Bryce, G. R. (2005). Developing tomorrow's statistician. *Journal of Statistics Education [online]*, *13*(1). URL: www.amstat.org/publications/jse/v13n1/bryce.html

Cobb G. W. (2007). The introductory statistics course: a ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*, 1-15.

Jones, B. & Sall, J. (2011). JMP statistical discovery software. *Wiley Interdisciplinary Reviews: Computational Statistics, 3*, 188–194. doi: 10.1002/wics.162

Lock, R., Lock, P. F., Lock, K. L., Lock, E. F., & Lock, D. F. (n.d.). StatKey to accompany *Statistics: Unlocking the Power of Data*. URL: http://lock5stat.com/statkey/

Pruim, R. J., Horton, N. J., & Kaplan, D. T. (2014). *Start Teaching with R*, *Preliminary Edition.* Project Mosaic.

Roseth, C. J., Garfield, J. B., & Ben-Zvi, D. (2008). Collaboration in learning and teaching statistics. *Journal of Statistics Education [online]*, 16(1). URL: http://www.amstat.org/publications/jse/v16n1/roseth.html

Rossman, A. J. & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics, 6*, 211–221. doi: 10.1002/wics.1302

Rossman, A. J., & Chance, B. L. (n.d.). Applets for *Introduction to Statistical Investigations*. URL: http://www.rossmanchance.com/ISIapplets.html

Spence, D. J., & Bailey, B. (2013). Technology-rich projects in statistics. In J. Foster (Ed.), *Proceedings of the 24th International Conference on Technology in Collegiate Mathematics, March 22-25, 2012* (pp. 173-177). Pearson Education, Inc.

Spence, D. J., & Bailey, B. (2015), Enhancing the benefits of discovery projects in elementary statistics. Paper presented at the AMS/MAA Joint Mathematics Meetings, San Antonio, TX, January 12, 2015.

Spence, D. J., Sharp, J. L., & Sinn, R. (2011). Investigation of factors mediating the effectiveness of authentic projects in the teaching of elementary statistics. *Journal of Mathematical Behavior, 30*, 319-332.