Paper 289-2010

Using New JMP® Interactive Modules to Teach Concepts in Introductory Statistics

Amy G. Froelich, Iowa State University, Ames, IA
William M. Duckworth, Creighton University, Omaha, NE

ABSTRACT

Simulation has become an important tool in teaching topics related to sampling distributions and inference in the introductory statistics class. Many of these simulations have been developed with Java applets and made available on the web. While these applets are easy to use and readily available to statistics instructors, they may not match classroom, laboratory or homework activities and are generally different than software used for data analysis. As a result, students can struggle with the transition between classroom activities and assignments, data analysis and conceptual understanding. In this paper we showcase new JMP Concept Discovery Modules developed using the JMP Scripting Language. These modules are designed to take advantage of JMP's powerful graphical interface features and to provide flexibility to match the simulation experience in JMP to classroom, laboratory or homework activities traditionally used in teaching concepts in introductory statistics.

INTRODUCTION

Simulation has become an important tool in teaching topics related to sampling distributions and inference in the introductory statistics class. Through the use of repeated sampling, simulation activities are designed to introduce students to concepts such as sampling variability, sampling distributions, coverage rates for confidence intervals, and Type I error rates for hypothesis tests. Many of these simulations have been developed with Java applets and made available on the web for instructor and student use. See the website by Rossman and Chance (2009) for a good example of available Java applets. Other simulations have been built into programs such as Fathom (Key Curriculum Press, 2009).

While these applets and programs are easy for instructors to use, they may not match the classroom, laboratory or homework activities used to introduce the concepts. These programs or Java applets are also generally different than software used for data analysis, with different terminology and graphics. As a result, students can struggle with the transition between classroom activities and assignments and between conducting data analysis and using computer simulation activities for conceptual understanding.

In this paper, we present new JMP Concept Discovery Modules for teaching concepts in the introductory statistics course. Programmed using the JMP Scripting Language, these modules were developed to solve the problems instructors and students find in using Java applets and other programs for building conceptual understanding in introductory statistics. The modules take advantage of the powerful graphical interface features used in JMP's data analysis platforms, thus bridging the gap between conceptual understanding and data analysis. Similar features appear in each module, providing the instructor and student with a set of unified tools for teaching and learning sampling distributions and inference. The modules are flexible and can be used to match the simulation experience with classroom, laboratory or homework activities. Finally, symbol notation is avoided to allow for maximum applicability to different textbooks and instructor's lecture notes.

In the following sections, the features of six of the new modules are described and a specific application of each module to a classroom activity or problem is presented. The modules for the sampling distribution for the sample proportion and the confidence interval and hypothesis test for the population proportion appear in Section 2. Similar modules for the mean appear in Section 3. Some ideas for future work are presented in Section 4 and conclusions are given in Section 5.

MODULES FOR THE PROPORTION

All modules have the same overall structure and format: user-supplied information on the left-hand side of the window, information from one generated sample on the upper right-hand side of the window and information collected from all generated samples on the lower right-hand side of the window. For the proportion modules, the user-supplied information includes the population proportion, the name of the category of interest, the sample size, and the number of samples to be generated. Animation can be turned on or off by radio button depending on the interest of the user.

For the sampling distribution module, users must choose to explore the center, spread or shape of the sampling distribution of the sample proportion by selecting the appropriate radio button. For the confidence interval module, the user must specify the confidence level for the calculated confidence intervals, either 80%, 90%, 95% or 99%. For the hypothesis testing module, the user must specify the alpha level of the hypothesis test, either 0.01, 0.05, or 0.1 and the direction of the alternative hypothesis, either less than, greater than or not equal to.

For each sample, category data are generated using JMP's built-in Binomial distribution data generation formula using the user-specified population proportion and sample size. Under animation mode, the bar graph and summary table for each generated sample appear in the upper-right hand side of each module. The confidence interval calculated from the sample appears below the bar graph and summary table in the confidence interval module and the hypothesis test calculated from the sample appears below the sample information in the hypothesis testing module. The output in this area of the module is similar to the output for a categorical variable from the JMP Distribution Analysis platform.

The lower right-hand side of the window is different for each module. For the sampling distribution module, the histogram of the sample proportions from all generated samples appears in this area. Depending on the radio button selected, the mean (Center) or standard deviation (Spread) of the generated sample proportions appears below the histogram. Selecting the Shape button adds a normal quantile plot above the histogram in the display. For the confidence interval module, a graph of all generated confidence intervals is given in this area along with a horizontal line at the user-supplied population proportion. A green line indicates the confidence interval includes the true population proportion and a red line indicates the confidence interval misses the true population proportion value. Only the most recent 100 calculated confidence intervals are visible, however, the JMP grabber tool can be used to view more confidence intervals. The proportion of all generated confidence intervals containing the true population proportion value is given below the graph. For the hypothesis testing module, the histogram of all generated z-test statistics appears in this area. A vertical red line denotes the rejection region of a one-sided test and two red vertical lines denote the same for a two-sided test. The proportion of all generated hypothesis tests rejecting the null hypothesis is displayed below the histogram.

In the three sub-sections below, an example of a classroom activity for each module is given, along with a description of how the module can be used to explore the important concepts in each activity.

SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

Before introducing inference for the population proportion, repeated sampling is used to present the sampling distribution of the sample proportion. Through simulation, students learn the concepts of sampling variability, the variability of the sample proportion statistic and the mean, standard deviation, and shape of the sampling distribution of the sample proportion. For example, based on data collected from students enrolled in a large introductory statistics course, approximately 35% of students reported having blue eyes. If samples of 200 students were selected from this population of students where 35% of students reported having blue eyes, how would the sample proportion of students with blue eyes vary from sample to sample?

In this example, the population proportion of 35% or 0.35, the category name, Blue Eyes and the sample size of 200 can be entered into the JMP module. Starting with one sample at a time allows students to see sampling variability through the changes in the bar graph and summary table in the upper-right-hand side of the window. Increasing to 10 or 100 samples at a time using animation begins the process of building the sampling distribution of the sample proportion in the histogram displayed in the lower-right-hand side of the window. For each additional generated sample, the histogram changes to incorporate the new sample proportion. Finally, the mean, standard deviation, and shape of the sampling distribution can be studied by simulating many, many samples (with or without animation). The results of 5000 repeated samples are depicted in the screen capture of the JMP Concept Discovery Module for the Sampling Distribution of the Sample Proportion in Figure 1.

Using the flexibility of the module, instructors can have students vary the population proportion, sample size and category name to match other activities and/or to investigate the success/failure condition for the sampling distribution of the sample proportion to be approximately normally distributed.

CONFIDENCE INTERVAL FOR THE POPULATION PROPORTION

The first topic in inference for the population proportion is generally the confidence interval. Repeated sampling is used to reinforce and connect the concept of sampling variability to the variability of the confidence interval and the connection between the coverage rate and confidence level of the confidence interval. The result is a visual representation of the concept of confidence and the interpretation of a confidence interval.

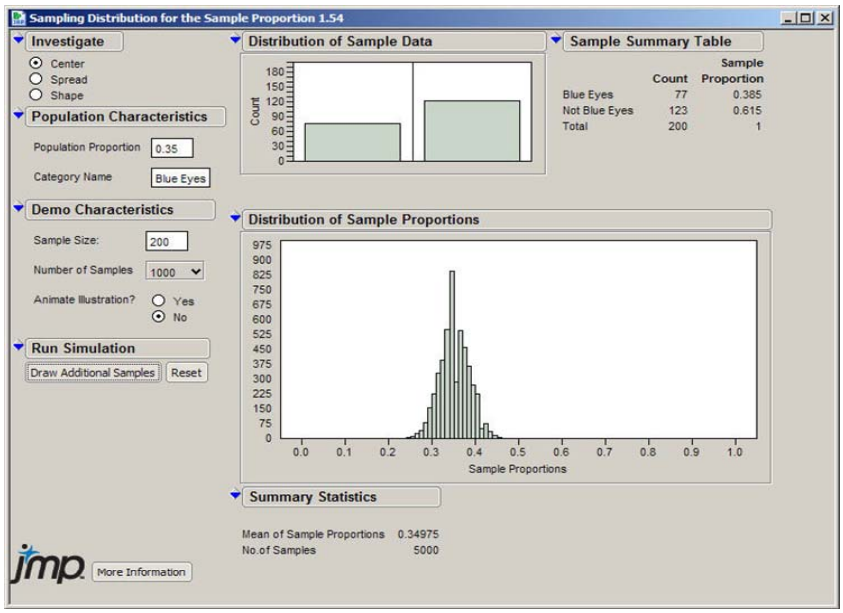

Figure 1: Sampling Distribution of the Sample Proportion Module

A typical activity to introduce confidence intervals for the population proportion is public opinion polling. For example, in a Gallop Poll taken June 16-18, 2009, 58% of 1504 national adults (aged 18 or older) surveyed stated they approved of the job performance of President Obama. The 95% confidence interval for the proportion of all U.S. adults who approve of the job performance of President Obama during this time frame is 55.47% to 60.45%. What does this confidence interval really mean?

In this example, the category of interest is the job approval of President Obama (Approve Obama) and the population proportion is the true proportion of U.S. adults who approve of the job performance of the president during this time frame. Since this value is unknown, we will assume a value of 60% or 0.6 for this proportion. The sample size is 1,504 U.S. adults and we will calculate 95% confidence intervals for this population proportion. Starting with one sample at a time allows students to see sampling variability through the changes in the bar graph and summary table in the upper-right-hand side of the window. They can also see the variability of the calculated confidence intervals. By generating 10 or 100 samples at a time, students see the variability of the confidence interval and the presence or absence of the true population proportion in each confidence interval in the confidence intervals graph. For many, many samples, students compare the observed percentage of confidence intervals containing the true population proportion value of 0.6 (the observed coverage rate) to the confidence level of 95%. The results of 100 repeated samples are depicted in the screen capture of the JMP Concept Discovery Module for the Confidence Interval for the Population Proportion in Figure 2.

Once a set of confidence intervals has been calculated, clicking on a particular confidence interval in the confidence interval graph displays the information about the corresponding sample and confidence interval under the Sample Confidence Interval heading. A different overall confidence level can also be selected, with the confidence interval graph and overall percentage of all confidence intervals that contain the population proportion dynamically changing with the new confidence level.

Using the flexibility of the module, instructors can have students vary the population proportion and sample size to match any activity and to investigate the effect of the success/failure condition on the coverage rate for the confidence intervals.

HYPOTHESIS TEST FOR THE POPULATION PROPORTION

The final module is the hypothesis test for the population proportion. Repeated sampling is used to reinforce and connect the concept of sampling variability to the variability of the z-test statistic and the Type I error rate of the hypothesis test. For example, the cracking rate of ingots used in manufacturing airplane parts is 20%. A new manufacturing process is designed to lower the proportion of cracked ingots. In a sample of 400 ingots manufactured with the new process, 72 cracked, for a sample proportion of 0.18. Is this evidence that the new process worked? What sample proportions would we expect to see if the cracking rate for the new process is the same as the old process?

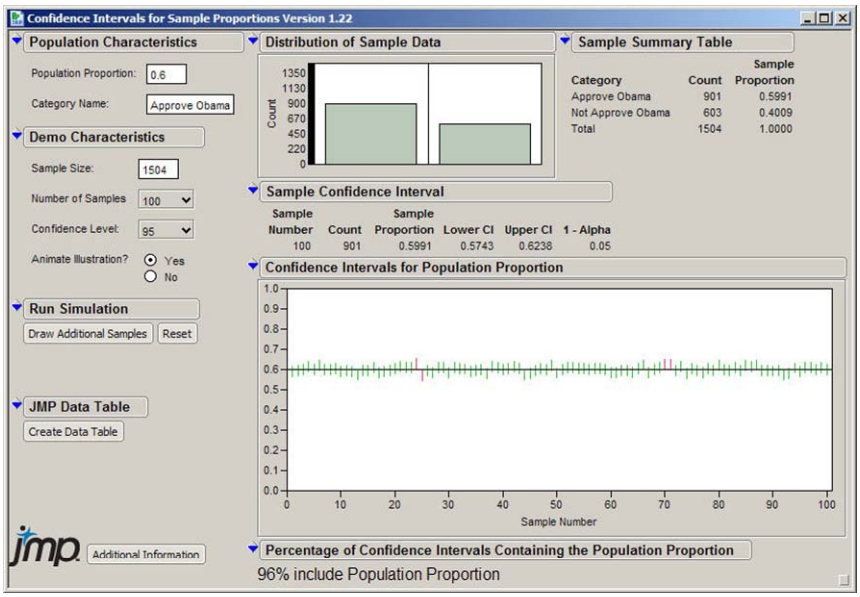

Figure 2: Confidence Interval for the Population Proportion Module

In this activity, the population proportion is 20% or 0.2, the category name is Cracked Ingot, and the sample size is 400. If the new process has worked, the proportion of cracked ingots should decrease making the alternative hypothesis Less Than. Since manufacturing processes are expensive to implement, we will require an alpha level of 0.05 for the hypothesis test.

Selecting one sample at a time shows students the sampling variability of the sample proportion and the variability of the z-test statistics calculated from the sample. Generating 10 or 100 samples at a time builds the variability and distribution of the z-test statistic from the hypothesis test while generating many, many samples allows students to compare the proportion of all hypothesis tests ending in a rejected null hypothesis (the observed Type I error) with the alpha level of the test. The results of 5000 repeated samples are depicted in the screen capture of the JMP Concept Discovery Module for the Hypothesis Test for the Population Proportion in Figure 3.

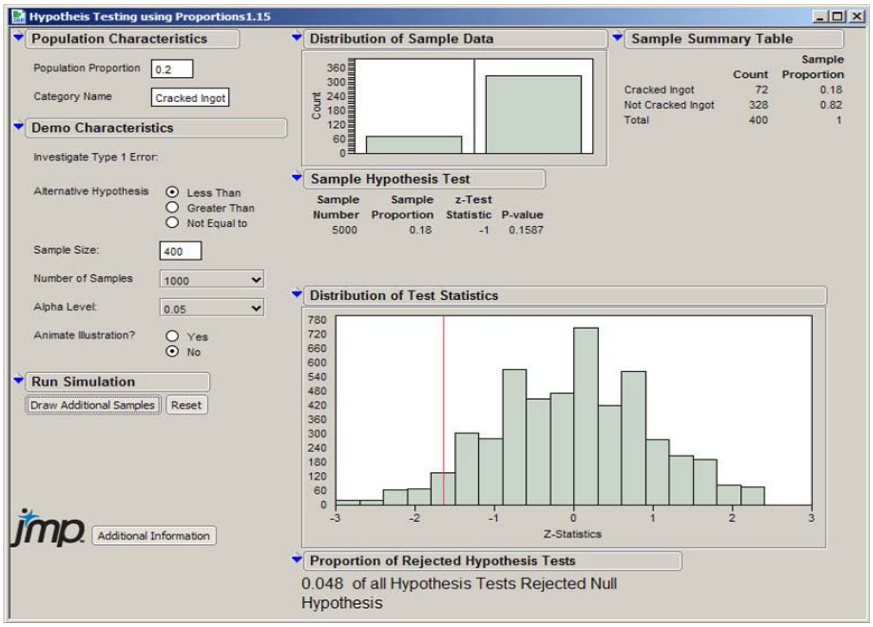

Figure 3: Hypothesis Test for the Population Proportion Module

Changing the alpha level after a set of hypothesis tests has been drawn dynamically changes the red line and the proportion of all rejected null hypothesis tests. Using the flexibility of the module, instructors can have students vary

the population proportion and sample size to match any activities and to investigate the effect of the success/failure condition on the distribution of the z-test statistics and the observed Type I error rate.

MODULES FOR THE MEAN

The format and structure of the sampling distribution, confidence interval and hypothesis testing modules for the mean are similar to the same modules for the proportion. On the left-hand side of the window, the user-supplied information includes the population distribution (Normal, Skewed, Uniform or My Data), the population mean and standard deviation, the variable name, the sample size, and the number of samples to be generated. Animation can be turned on or off by radio button depending on the interest of the user. For the sampling distribution module, users can choose to explore the center, spread or shape of the sampling distribution of the sample mean by selecting the appropriate radio button. For the confidence interval module, the user must specify the confidence level for the calculated confidence intervals, either 80%, 90%, 95% or 99%. For the hypothesis testing module, the user must specify the alpha level of the hypothesis test, either 0.01, 0.05, or 0.1 and the direction of the alternative hypothesis, either less than, greater than or not equal to. For both the confidence interval and hypothesis testing modules, users must specify whether or not to use the known population standard deviation for inference.

For each sample, values are generated based on the user-specified distribution, either Normal, Uniform, Skewed, or My Data. For the Normal distribution option, the population mean and standard deviation are used as the data generation parameters for JMP's built-in data generation formula for the Normal distribution. For the Uniform distribution, the population mean and standard deviation are used to transform values generated using JMP's built-in Uniform distribution formula. For the Skewed distribution option, values are generated from JMP's built-in Exponential distribution formula with mean 1 and then transformed to have the user-specified mean and standard deviation. In all cases, the data are generated as independent and identically distributed from the given distribution.

For the My Data option, users have the ability to use their own population as the basis of the simulation. This population should be large (more than 10 times the desired sample size) and be contained as a column in a JMP data file. Each sample is selected without replacement from this large population. The values of the population mean and standard deviation are not entered by the user for this option, but calculated directly from the population data. Selecting a sample size larger than 10% of the population size will result in an error message.

Under animation mode, the histogram and summary table for each generated sample appear in the upper-right hand side of each module. The confidence interval calculated from the sample appears below the histogram and summary table in the confidence interval module and the hypothesis test calculated from the sample appears below the sample information in the hypothesis testing module. If the user selected to use the known population standard deviation for inference, a z-distribution based confidence interval or hypothesis test will be calculated. Otherwise, the t-distribution based confidence interval or hypothesis test will appear. The output in this area of the module is similar to the output for a continuous variable from the JMP Distribution Analysis platform.

The lower right-hand side of the window is different for each module. For the sampling distribution module, the histogram of the sample means from all generated samples appears in this area. Depending on the radio button selected, the mean (Center) or standard deviation (Spread) of the generated sample means appears below the histogram. Selecting the Shape button adds a normal quantile plot above the histogram in the display. For the confidence interval module, a graph of all generated confidence intervals is given in this area along with a horizontal line at the user-supplied population mean. Again, green lines indicate the confidence interval includes the true population mean and red lines indicate the confidence interval misses the true population mean value. Only the most recent 100 calculated confidence intervals are visible, however, the JMP grabber tool can be used to view more confidence intervals. The proportion of all generated confidence intervals containing the true population mean value is given below the graph. For the hypothesis testing module, the histogram of all generated test statistics appears in this area. A vertical red line denotes the rejection region of a one-sided test and two red vertical lines denote the same for a two-sided test. The proportion of all generated hypothesis tests rejecting the null hypothesis is displayed below the histogram.

In the three sub-sections below, an example of a classroom activity for each module is given, along with a description of how the module can be used to explore the important concepts in each activity.

SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Before introducing inference for the population mean, repeated sampling is used to present the sampling distribution of the sample mean. Through simulation, students learn the concepts of sampling variability, the variability of the sample mean statistic and the mean, standard deviation, and shape of the sampling distribution of the sample mean.

Selecting a non-normal population distribution also allows students to explore the Central Limit Theorem for the sample mean.

For example, the mean height of adult females in the US population is 63.8 inches with a standard deviation of 2.8 inches. How will the mean height of samples of 200 females from this population vary? Starting with one sample at a time allows students to see sampling variability through the changes in the histogram and summary table in the upper right-hand side of the window. Increasing to 10 or 100 samples at a time using animation begins the process of building the sampling distribution of the sample mean in the histogram displayed in the lower right-hand side of the window. For each additional generated sample, the histogram changes to incorporate the new sample mean. Finally, the mean, standard deviation, and shape of the sampling distribution can be studied by simulating many, many samples (with or without animation). The results of 5000 repeated samples are depicted in the screen capture of the JMP Concept Discovery Module for the Sampling Distribution of the Sample Mean in Figure 4.

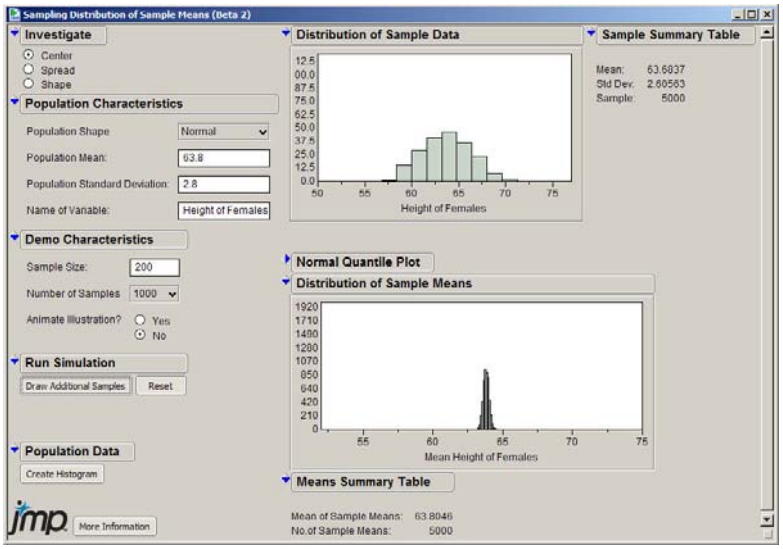

Figure 4: Sampling Distribution for the Sample Mean Module

Using the flexibility of the module, instructors can have students vary the population distribution, mean and standard deviation, sample size and category name to match other activities and/or to investigate the sample sizes needed for the sampling distribution of the sample mean to be approximately normally distributed for different non-normal population distributions.

CONFIDENCE INTERVAL FOR THE POPULATION MEAN

The first topic in inference for the population mean is generally the confidence interval. Repeated sampling is used to reinforce and connect the concept of sampling variability to the variability of the confidence interval and the connection between the coverage rate and confidence level of the confidence interval. The result is a visual representation of the concept of confidence and the interpretation of a confidence interval.

For example, in a random sample of 52 adults, the sample mean body temperature was 98.28 degrees with a sample standard deviation of 0.68 degrees. The 95% confidence interval for the mean body temperature of the population is 98.09 to 98.47 degrees. Medical literature generally gives the mean body temperature of adults as 98.6 degrees, which is not in our confidence interval. If the mean body temperature of adults was really 98.6 degrees, what would the confidence intervals from a random sample of 52 adults from this population look like?

Starting with one sample at a time allows students to see sampling variability through the changes in the histogram and summary table in the upper right-hand side of the window. They can also see the variability of the calculated confidence intervals. Some students think the confidence interval gives an interval for values from the sample instead of values for the population mean. Comparing the histogram and summary statistics for several samples to the calculated confidence interval from the same sample can help to fix this misconception. Generating 10 or 100 samples at a time helps students see the variability of the confidence interval and the presence or absence of the true population mean in each confidence interval in the confidence intervals graph. For many, many samples, students compare the observed percentage of confidence intervals containing the true population mean value (the observed

coverage rate) to the confidence level of 95%. The results of 500 repeated samples are depicted in the screen capture of the JMP Concept Discovery Module for the Confidence Interval for the Population Mean in Figure 5.

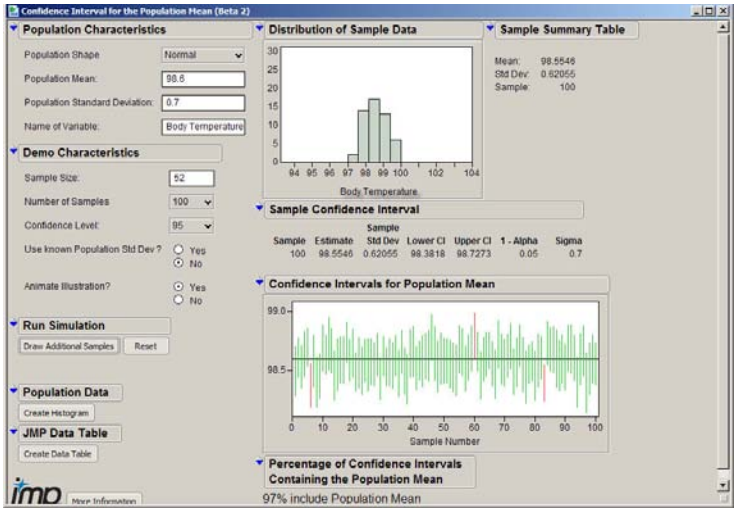

Figure 5: Confidence Interval for the Population Mean Module

Once a set of confidence intervals has been calculated, clicking on a particular confidence interval in the confidence interval graph displays the information about the corresponding sample and confidence interval under the Sample Confidence Interval heading. Students can see that if the mean body temperature of this population were really 98.6 degrees, 95% of all confidence intervals would contain that value and very few confidence intervals would be as low as the one obtained from our given sample. A different overall confidence level can also be selected, with the confidence interval graph and overall percentage of all confidence intervals that contain the population mean dynamically changing with the new confidence level.

Using the flexibility of the module, instructors can have students vary the population distribution, mean, and standard deviation and the sample size to match any activity and to investigate the effect changes for each component have on the observed coverage rate for the confidence intervals.

HYPOTHESIS TEST FOR THE POPULATION MEAN

The final module is the hypothesis test for the population mean. Repeated sampling is used to reinforce and connect the concept of sampling variability to the variability of the z-test or t-test statistic and the Type I error rate of the hypothesis test. For example, the SAT Math test is known to have a population mean score of 500 points with a standard deviation of 100 points. The 100 students in my introductory statistics course have a mean SAT Math score of 510 points. Are these students like a random sample of students from this population with respect to their mean SAT Math Score? Or is their mean score larger than what would be expected from a random sample of 100 students from this population?

Selecting one sample at a time shows students the sampling variability of the sample mean and the variability of the z-test statistic (in this example) calculated from the sample. Generating 10 or 100 samples at a time builds the variability and distribution of the z-test or t-test statistic from the hypothesis test while generating many, many samples allows students to compare the proportion of all hypothesis tests ending in a rejected null hypothesis (the observed Type I error) with the alpha level of the test. Students see that the mean SAT Math score from the introductory statistics course of 100 students has a similar mean score as a random sample of 100 students from this population. The results of 500 repeated samples are depicted in the screen capture of the JMP Concept Discovery Module for the Hypothesis Test for the Population Proportion in Figure 6.

Changing the alpha level after a set of hypothesis tests has been drawn dynamically changes the red line and the proportion of all rejected null hypothesis tests. Using the flexibility of the module, instructors can have students vary the population distribution, mean, and standard deviation and sample size to match any activities and to investigate the effect of changes in these components on the distribution of the test statistics and the observed Type I error rate.

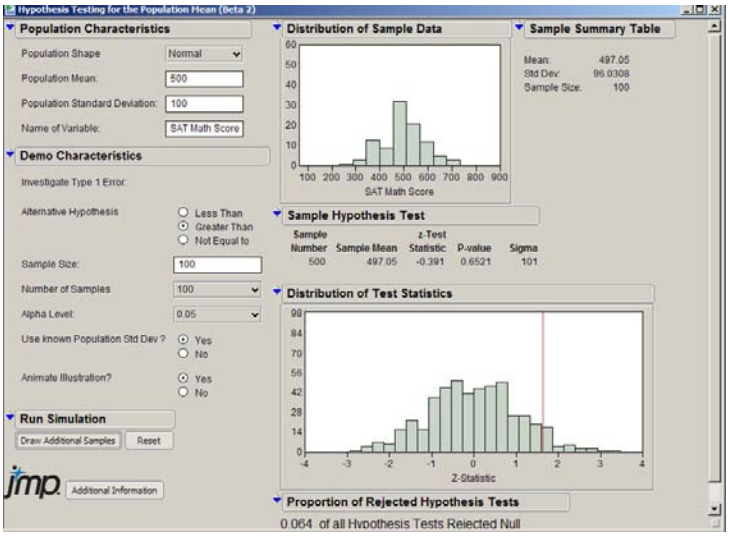

Figure 6: Hypothesis Test for the Population Mean Module

FUTURE WORK

The six modules described in Sections 2 and 3 of this article are part of an overall collection of modules for teaching and learning concepts in introductory statistics. Additional modules either completed or in production include ones for teaching concepts for the distribution of a quantitative variable, regression and one-way ANOVA. The formats for all modules are similar and provide the user with a unified approach to learning and teaching concepts in introductory statistics.

The ultimate goal for the use of the modules is to provide a better learning experience for students and to increase understanding of the important concepts in introductory statistics courses. We are currently writing additional examples for using the modules for classroom, laboratory and homework activities. With these activities, we will then test whether the use of these modules leads to better understanding of the concepts in introductory statistics.

CONCLUSIONS

Through repeated sampling, simulation activities are designed to introduce students to concepts in sampling and inference. While many Java applets are available for these simulations, students can struggle with the transition between classroom activities and assignments, data analysis and computer simulation activities. The JMP Concept Discovery Modules in this paper were designed to solve these problems. They are flexible: they can be used to match any classroom or homework activities. Symbol notation is avoided, making these modules applicable to any textbook or instructors notes. The interactive graphics and output in the modules are the same as the ones used in the JMP data analysis platforms, making the transition between conceptual understanding and data analysis seamless.

ACKNOWLEDGEMENTS

The JMP Concept Discovery Modules in this paper were developed by the two authors and programmed by Predictum, Inc. Funding for development and programming was provided by JMP Statistical Discovery Software, a business division of SAS Institute, Inc. All modules, including the ones described in this paper, are available for download free of charge at the JMP website at http://www.jmp.com/academic/learning_modules.shtml.

REFERENCES

Key Curriculum Press (2009). Fathom Dynamic Data Software. Version 2.11.

Rossman, A. & Chance, B. (2009). Rossman/Chance Applet Collection. www.rossmanchance.com/applets.html.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the lead author at:

Amy G. Froelich
Iowa State University
3109 Snedecor Hall
Ames, IA 50011-1210
Phone: 515-294-5584
Fax: 515-294-4040
E-mail: amyf@iastate.edu
Web: http://www.public.iastate.edu/~amyf/homepage.html