USING INTERACTIVE SIMULATION MODULES IN JMP® FOR SELF-PACED LEARNING OF INTRODUCTORY STATISTICS

David Meintrup and Volker Kraft
University of Applied Sciences Ingolstadt, Germany
JMP Academic Team, Germany
David.Meintrup@thi.de

An excellent way to deepen students' understanding of statistical data analysis is to first have them perform experiments and then analyze the corresponding data. While this approach is already time-consuming and labor-intensive under normal circumstances, it becomes unfeasible in times of distance learning. Therefore, we replaced actual experimentation by interactive simulation modules. The statistical software JMP® includes 10 interactive teaching modules, covering basic concepts of statistics. We used these modules for a course in Engineering Statistics at a German University. The modules were well-received, students particularly appreciated three aspects: the ease of use, the interactivity of the modules, and the self-paced and time-independent form. In summary, JMP's interactive simulation modules were helpful in deepening the understanding of basic statistical methods for engineering students. They can easily be integrated in introductory statistics classes as a tool for improving statistical literacy, for both distance learning or in-class learning situations.

INTRODUCTION

The German word for "understand" is "begreifen". The root of this word literally means to "grab something with your hands". It demonstrates the close connection between the physical experience of constructing something and the mental process of understanding its meaning. Anyone who has ever taught a class on Design of Experiments (DoE) using Box' famous paper helicopter experiment (Box, 1991) can confirm: once the students built several helicopters, measured their flight times and analyzed the data, they made a huge step towards understanding what DoE is all about.

There can be several reasons why performing actual experiments will not be feasible. In large groups, it might simply be impossible to organize, it is very time-consuming, and of course, it requires physical presence in classrooms, which has been impossible for more than a year now due to the coronavirus pandemic.

The closest we can come to performing experiments is simulating them. Of course, computer-simulated experiments do not give you the same "touch-and-feel" as real experiments, but they come with some advantages: it is very easy to change prerequisites and get a whole new set of experiments, they can be repeated as often as necessary at no costs, and they can be performed by each student individually – in a classroom or at home. Ideally, the simulations should allow the user to interactively change parameters and see how this impacts the outcome of the analysis.

The statistical software package JMP® (JMP, 2021) comes with 10 interactive simulation modules that are aimed to ease the understanding of some fundamental concepts of statistics. All of them allow the user to interactively change setting, draw additional samples and re-analyze the data. Table 1 contains an overview of the 10 simulation modules with a short description of their purpose.

We used these teaching modules in a course in Engineering Statistics at the University of Applied Sciences in Ingolstadt, Germany. The students who attended this class had one year of engineering mathematics, but no prior knowledge of statistics. The goal of the course was to introduce fundamental concepts of applied statistics, typically including descriptive statistics, visualizations, estimators, hypothesis testing, simple regression and one-way analysis of variance. The emphasis lied on practical data analysis rather than theory, and included using JMP® as enabling statistical software package.

The teaching modules were very helpful in order to help students understanding fundamental concepts of statistics. Therefore, we will present two out of the 10 in more detail, and discuss our experience using them for Engineering Statistics.

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

Table 1. Interactive Teaching Modules

No.	Teaching Module	Description
(1)	Distribution Generator	Displays the distribution and descriptive statistics of a given dataset
(2)	Sampling Distribution of Sample Means	Draws samples from different distributions (normal, skewed, uniform) and shows the distribution of the sample means
(3)	Sampling Distribution of Sample Proportions	Draws samples from a binomial distribution and shows the distribution of the sample proportions
(4)	Confidence Interval for Population Mean	Draws samples from different distributions (normal, skewed, uniform) and shows the corresponding confidence intervals for the population mean
(5)	Confidence Interval for Population Proportion	Draws samples from a binomial distribution and shows the corresponding confidence intervals for the population proportion
(6)	Hypothesis Test for Population Mean	Draws samples from different distributions (normal, skewed, uniform) and shows the distribution of the test statistic for testing the mean
(7)	Hypothesis Test for the Population Proportion	Draws samples from a binomial distribution and shows the distribution of the test statistic for testing the mean
(8)	Distribution Calculator	Shows the density function of 25 different distributions with adjustable parameters and calculates quantiles and probabilities
(9)	Demonstrate Regression	Draws correlated samples following a linear model and displays the corresponding linear fit
(10)	Demonstrate ANOVA	Draws samples of two or three groups with given group means and displays the distribution of the test statistic and the analysis of variance

INTERACTIVE TEACHING MODULES

The general structure is the same for most of the 10 interactive teaching modules listed in Table 1, which simplifies their use. On the left side of the report window (see Figure 1 for an example), one can specify the parameters and characteristics of the population the data is drawn from. The sample size and the number of samples can be specified, and the button “Draw Additional Samples” starts the simulation process. In the middle part, graphical representations of the data and the analysis are given. Finally, the right part of the report window contains statistical values relevant to the performed analysis. In the following, we will present two simulation modules in more detail, one complex one (“Demonstrate Regression”) and one simpler module (“Confidence Interval for the Population Mean”).

The module “Demonstrate Regression” explains the concept of a simple linear regression. It has three columns, as can be seen in the screenshot in Figure 1. In the left column, the user can start by specifying the source of the data. In the shown example, the x-variable is normally distributed, while the y-variable is simulated using the model $y = 5 + 3x$. The user can change the intercept, the slope and the degree of correlation between x and y. The x- and y-variable can be renamed, and the mean and standard deviation can be assigned. Below, a sample size (here: 25) and the number of drawn samples (here: 100) can be specified. Each time that the user hits the button “Draw Additional Samples” a new set of samples is generated using the determined population, and all graphs and statistics are updated.

On the top of the middle column, the x- and y-variable are displayed in a scatterplot. The user can interactively fit a line to the data, display the squares of the residuals (blue squares), and compare the sum of squared residuals to the best fit (red line - SSE). The model function is given for the last sample. Below is a graph showing all 100 fitted linear regression models for the 100 samples.

On the right side, the summary of fit contains the R^2 value, the root mean squared error (RMSE), the mean response value and the number of observations. The analysis of variance tests if the slope parameter is significantly different from 0, just as the parameter test in the "Parameter Estimates" section does, using a t-test. Finally, the residual-by-predicted plot is presented for diagnostic purposes.

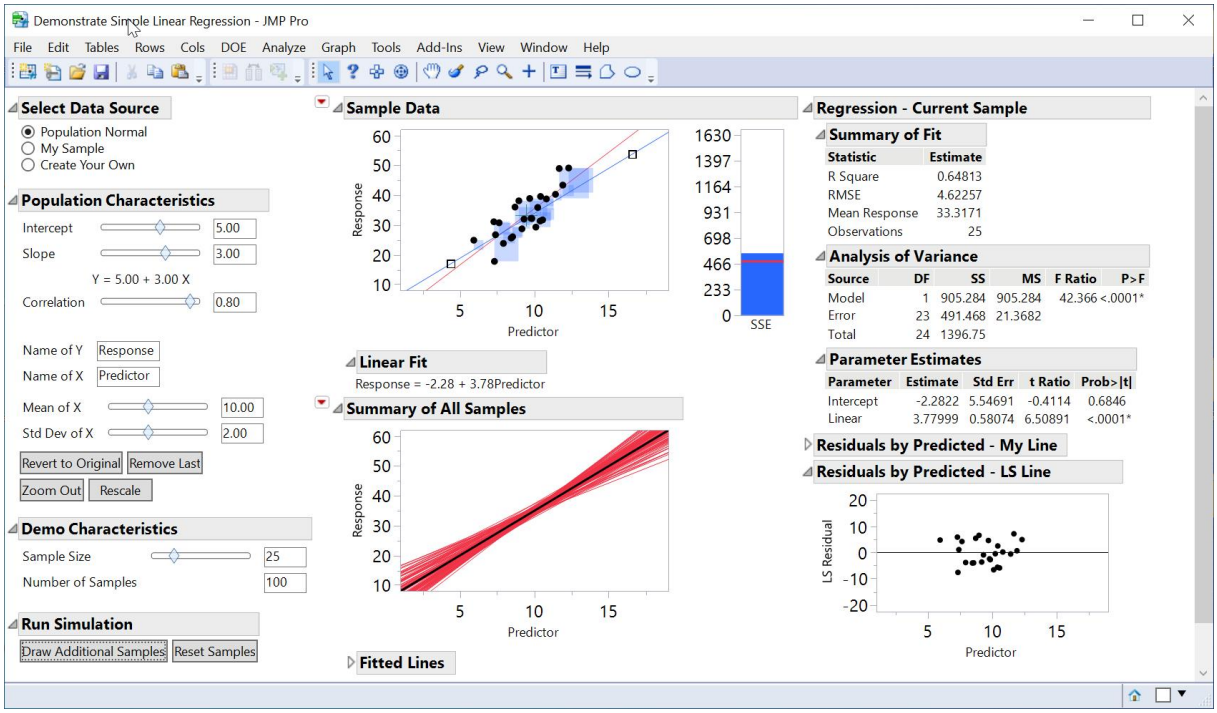

Figure 1. The interactive simulation module "Demonstrate Regression"

Despite the simplicity of a simple linear regression model, this module can teach three fundamental concepts that are used much more generally in statistical modeling (Harrell, 2016). The first idea is that every supervised learning technique needs a loss function. In this case, the sum of squared residuals is used as a loss function, and its value is visualized by the blue bar on the right side of the scatterplot. Moving the regression line, the student can observe how the value of the loss function changes, until it reaches its minimum once the chosen line corresponds to the best fit.

The second fundamental idea is the presence of uncertainty in any statistical modeling. In the presence of noisy data, different samples will lead to different linear models. Typically, this uncertainty is reflected by presenting the fitted regression line with a confidence band and the parameters of the fit with confidence intervals. As useful as we believe these statistics are, a more direct impression of uncertainty can be gained by fitting 100 lines and drawing them in one plot (Figure 1 – Summary of All Samples, red lines). Like this, the user can observe by himself how the confidence band around the linear fit emerges. In particular, one gets a different perspective why the confidence band is larger further away from the center of the data set.

Finally, every statistical model has some assumptions, and checking their validity should be part of each modeling process. The residual-by-predicted plot (Figure 1 – right column) allows to check for outliers, and to assess homoscedasticity. Obviously, there is more to do for a thorough check of regression assumptions, but it is a good starting point to get people used to not skipping residual diagnostics when fitting statistical models.

The teaching module on linear regression – as some other modules as well – allows to import your own data set instead of using simulated data. This is particularly useful for homework assignments, as questions can be targeted to the specific data set.

The interactive teaching module "Confidence Interval for the Population Mean" is shown in Figure 2. On the left side, the user can specify the parameters of the population, the size and number of samples, and the confidence level. In the middle, the top figure displays a single sample, while the bottom graph shows all the confidence intervals for the 100 samples that have been drawn. Below, summary statistics for the 100 samples are given. The display on the right side shows the mean and standard deviation of the last sample.

For novices in statistics, the false idea of a 95% confidence meaning "there is a 95% chance that the confidence interval contains the true mean" is very attractive (Greenland et al., 2016), and convincing them that this interpretation is wrong, is difficult. The module makes it very easy to convey the true idea of a confidence interval: if repeated many times, on average 95% of the confidence intervals will contain the true value, and 5% will not. The graph in the middle shows the confidence intervals that do not contain the true mean in red, the others in green. In the shown example, 95 of 100 confidence intervals do contain the true mean (which, of course, is not always the case), and this is precisely the right interpretation of a 95% confidence interval.

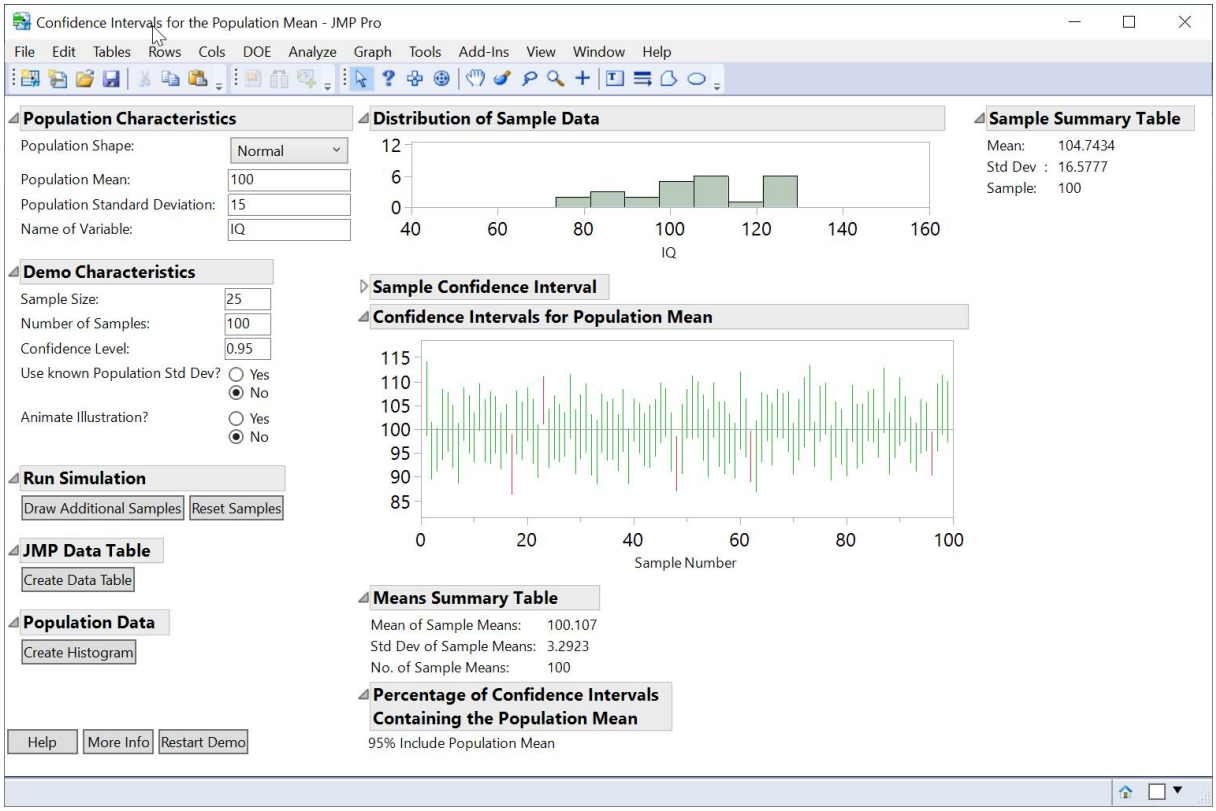

Figure 2. The interactive simulation module "Confidence Interval for the Population Mean"

DISCUSSION

We used all 10 interactive teaching modules in our introductory class on Engineering statistics. In the fall semester of 2020, no in-class lectures were allowed at the University of Applied Sciences in Ingolstadt. Therefore, the course was exclusively taught online, and consisted of one lecture per week, homework assignments and a final exam. Typically, the topic (e.g., simple linear regression) was introduced during the lecture. Using the simulation module to deepen the understanding of the corresponding topic was part of the homework. Whenever possible, we provided an example data set and asked some questions about it that could be answered with the help of the simulation module.

Besides the just described classic approach to integrate the modules into homework assignments, we would like to mention three other didactical concepts that can be supported by the interactive teaching modules:

- Peer-to-peer teaching (Stigmar, 2016): one subgroup studies, for example, the use of the linear regression module and explains its functionality to the other students.
- Flipped classroom (Colomo-Magaña et al., 2020): in particular in the presence of video tutorials, the modules can be very helpful for students when they are supposed to learn the corresponding concept at home, as in a flipped classroom situation.
- Gamification (Legaki et al, 2020): asking questions that can be answered with the modules in class and then giving points for correct answers, converts learning about regression into a game.

We see several advantages in using these or similar simulation tools in a statistics class. The students can use them at their own pace, as often and long as they need or want. They can use them again at a later time point when they are preparing for the exam. The modules are very easy to use, the students don't need any explanation. Exploring the functionality of the modules can be seen as part of the statistical learning process. The modules can very easily be integrated into homework assignments with given data sets. Finally, once the software package is available, there are no extra costs involved. The statistical software JMP® is a very interactive in its regular use, so that the modules can be integrated seamlessly.

We don't want to claim that the use of the teaching modules improved the results of the students in the final exam, as this would require a more carefully designed experiment, including a control group. However, the students' feedback was very positive. They particularly appreciated the interactivity, the ease of use, and the self-paced and time-independent form. In summary, we believe that the interactive simulation modules helped deepening the understanding of basic statistical methods for engineering students.

REFERENCES

Box, G.E.P. (1991). Teaching Engineers Experimental Design with a Paper Helicopter. *Report no. 76, Center for Quality and Productivity Improvement.* University of Wisconsin.

Colomo-Magaña, E., Soto-Varela, R., Ruiz-Palmero, J., Gómez-García, M. (2020). University Students' Perception of the Usefulness of the Flipped Classroom Methodology. *Education Sciences*, *10*, 275.

Greenland, S., Senn, S.J., Rothman, K.J. et al. (2016). Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology 31*, 337–350.

Harrell, F. E., Jr. (2016). Regression modeling strategies. *Springer International Publishing*. Cham, Switzerland.

JMP® (2021). *Version 15.2.0, SAS Institute Inc.*, Cary, NC, 1989-2021. *www.jmp.com*

Legaki, N. Z., Xi, N., Hamari, J., Karpouzis, K., & Assimakopoulos, V. (2020). The effect of challenge-based gamification on learning: An experiment in the context of statistics education. *International Journal of Human-Computer Studies, 144.*

Stigmar, M. (2016). Peer-to-peer Teaching in Higher Education: A Critical Literature Review. *Mentoring & Tutoring: Partnership in Learning, 24*(2).