Fake Data, Real Learning

A new R package to use synthesized data in the classroom

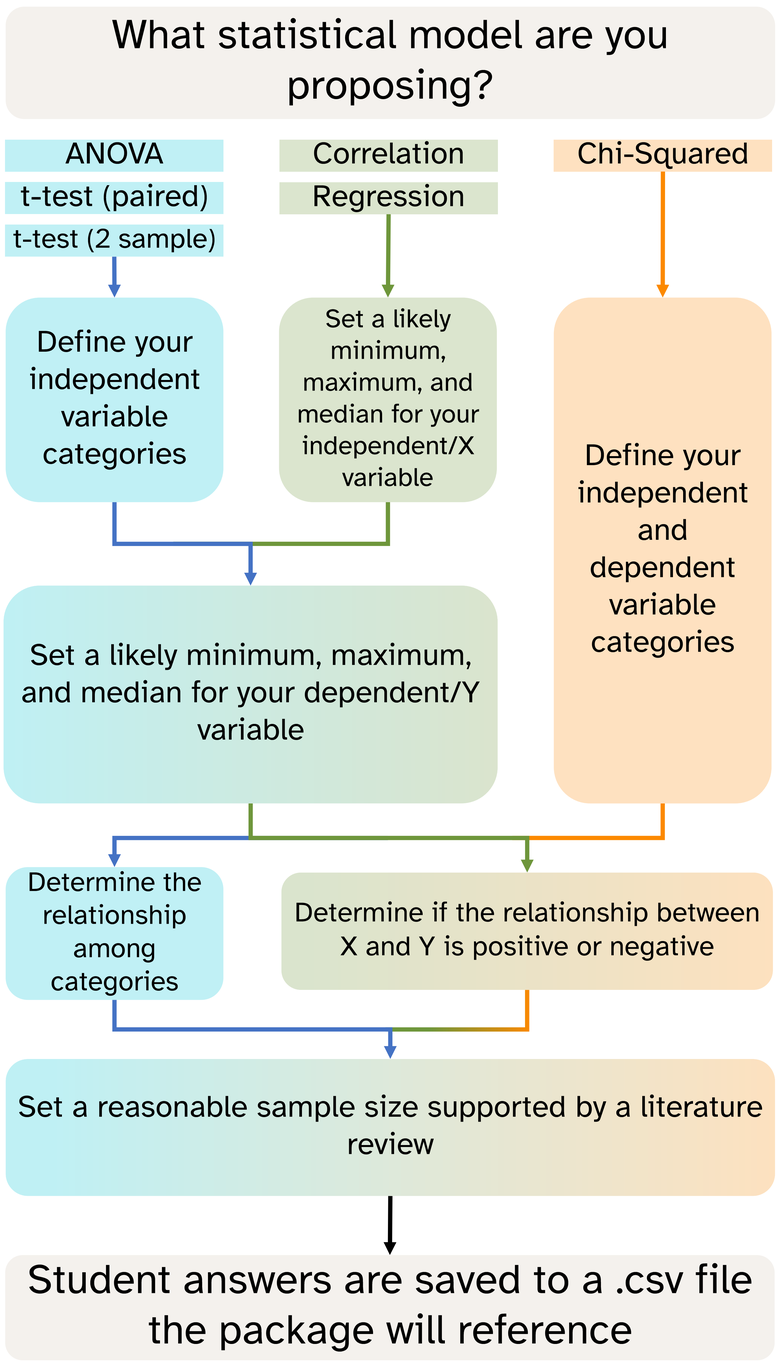
Liam O. Mueller and Keefe Reuther
University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093

Problem: Given the structure of most college courses, it can be difficult to have students generate an experimental design, collect data, and analyze those data in a 10-16 week course. Work arounds for many classes have included giving students a predetermined experimental design, or having the analysis pre-described. The most time consuming step in courses built around teaching the steps of the scientific cycle is the collection of real world data.

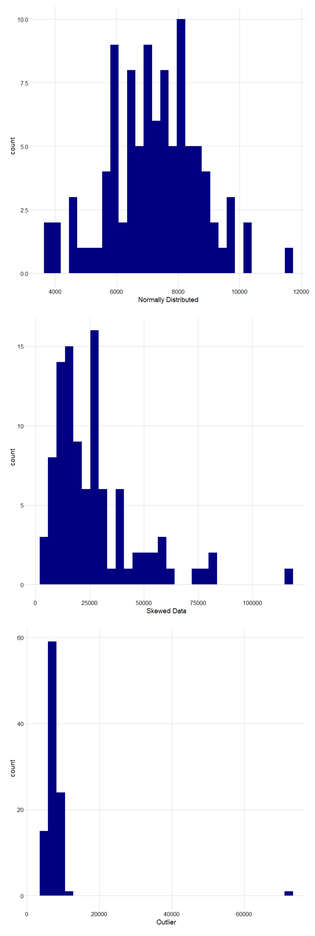
Solution: The "DataSimulatoR" package generates unique, lifelike data for each student based on a series of simple inputs. For instructors, it provides grading keys including figures and analysis summaries. The package can generate data suitable for univariate analyses common in undergraduate education including: linear models with a single X and Y variable (referred to as "ANOVA" & "Regression"), two sample t-tests, paired t-tests, correlation, and chi-square (Categorical X & Y analysis).

How it Works

Step 1: Students request data by answering a series of simple questions about their experimental design using a Google form.

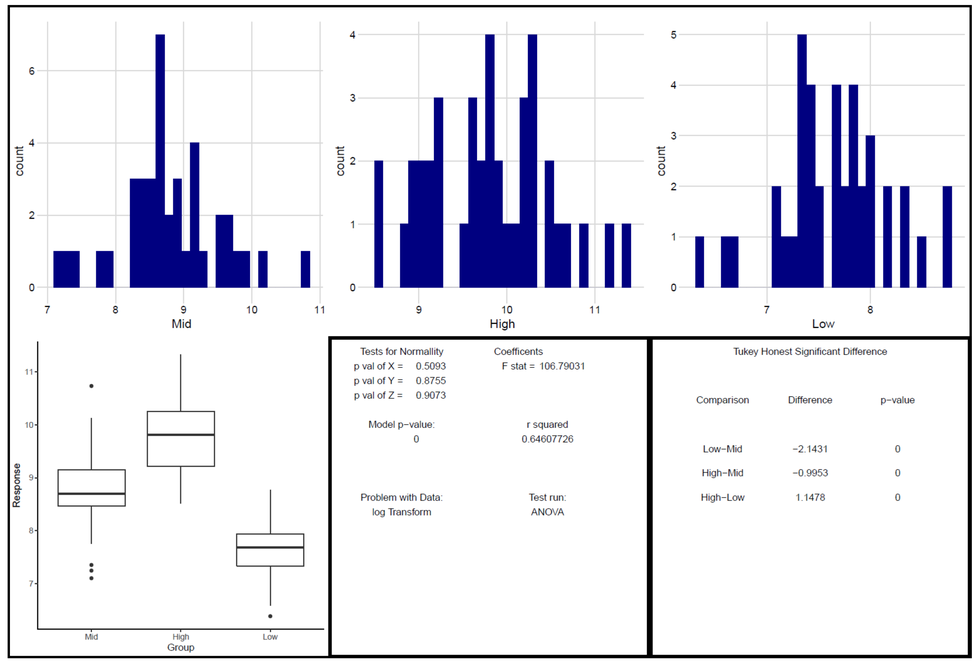


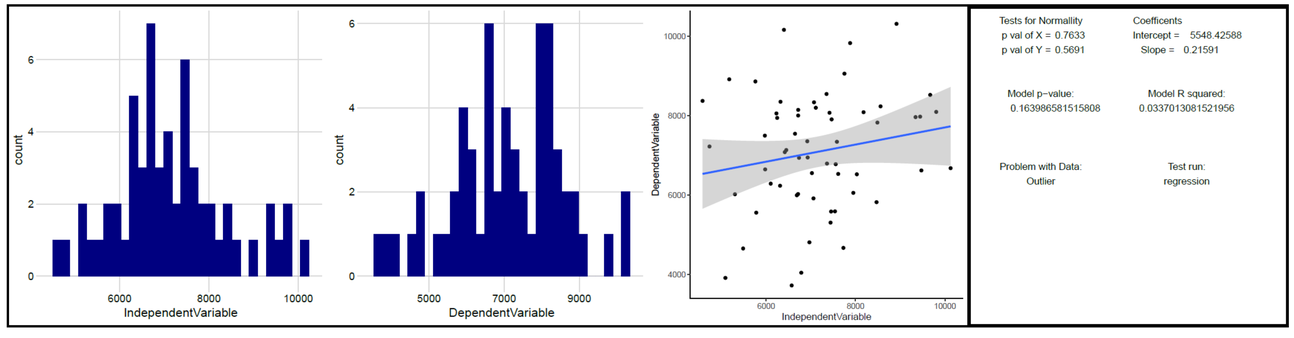
Step 2: Instructors set parameters for the simulator including the addition of "lifelike" problems in the data like skew and outliers.



Step 3: The R code generates .csv files unique for each student and emails them the file. The instructor receives these files and a grading key.

Keys created by the package are shown for a one way ANOVA (right) and linear regression (below). These keys include information about the instructor parameters and expected student analyses, including tests for normality, model coefficients, and p-values.





Try it Yourself

Follow the QR code to a GitHub repository containing:

- The DataSimulatoR code
- Links to the google forms
- README file
- License: GNU_GPL 3.0
- A feedback form. Please tell us what you would like this R package to do for you!

Observations and Thoughts on Implementation

- **See Dr. Keefe Reuther's poster today here at SABER West for a full discussion on the success of implementing student projects with the "DataSimulatoR" package.**

Package Dependencies and Citations

This package was written for the R programming language[1], using RStudio[2]. This package relies upon the following three dependencies: cowplot[3], rockchalk[4], and tidyverse[5].

[1]R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[2]RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
[3]Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot.
[4]Paul E. Johnson (2022). rockchalk: Regression Estimation and Presentation. R package version 1.8.157. https://CRAN.R-project.org/package=rockchalk.
[5]Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 1686. doi: 10.21105/joss.01686 (URL: https://doi.org/10.21105/joss.01686).