

Randomization-Based Statistical Inference: A Resampling and Simulation Infrastructure

Ivo D. Dinov^{1,2,3}, Selvam Palanimalai¹, Ashwini Khare¹, and Nicolas Christou¹

¹ Statistics Online Computational Resource, University of California, Los Angeles, Los Angeles, CA 90095

² Statistics Online Computational Resource, University of Michigan, UMSN, Ann Arbor, Michigan 48109-5482

³ Michigan Institute for Data Science, University of Michigan, Ann Arbor, Michigan 48109

Corresponding Author:

Ivo D. Dinov, Director

Statistics Online Computational Resource

University of Michigan, UMSN

426 North Ingalls, Suite 4100

Ann Arbor, Michigan 48109-5482

Tel: (734) 615-5087

Fax: (734) 647-2416

www.SOCR.umich.edu

statistics@umich.edu

Abstract

Statistical inference involves drawing scientifically-based conclusions describing natural processes or observable phenomena from datasets with intrinsic random variation. There are parametric and non-parametric approaches for studying the data or sampling distributions, yet few resources are available to provide integrated views of data (observed or simulated), theoretical concepts, computational mechanisms and hands-on utilization via flexible graphical user interfaces.

We designed, implemented and validated a new portable randomization-based statistical inference infrastructure (http://socr.umich.edu/HTML5/Resampling_Webapp) that blends research-driven data analytics and interactive learning, and provides a backend computational library for managing large amounts of simulated or user-provided data. The core of this framework is a modern randomization webapp, which may be invoked on any device supporting a JavaScript-enabled web-browser. We demonstrate the use of these resources to analyze proportion, mean, and other statistics using simulated (virtual experiments) and observed (e.g., Acute Myocardial Infarction, Job Rankings) data. Finally, we draw parallels between parametric inference methods and their distribution-free alternatives.

The Randomization and Resampling webapp can be used for data analytics, as well as for formal, in-class and informal, out-of-the-classroom learning and teaching of different scientific concepts. Such concepts include sampling, random variation, computational statistical inference and data-driven analytics. The entire scientific community may utilize, test, expand, modify or embed these resources (data, source-code, learning activity, webapp) without any restrictions.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1111/test.12156](https://doi.org/10.1111/test.12156)

Keywords: resampling, simulation, statistical inference, randomization, bootstrapping, Statistics Online Computational Resource (SOCR).

Author Manuscript

I. Introduction

The core of *statistical inference*, the process of drawing data-driven conclusions and decision-making, is based on the concepts of *random sampling* and *sampling distributions*. A sample is an observed collection of data chosen from a specific population of interest (or from the same probability distribution). A sample is random if it is chosen by a method involving an unpredictable stochastic component where sequential data are independently observed. It is a common assumption that random samples are representative of the population (or distribution) from which they are drawn. Otherwise, discrepancies between the sample and the distribution of the natural process may introduce a *sampling error*, which could negatively affect the statistical inference, even if we can estimate the magnitude of the sampling error (Lohr, 2009). A *statistic* is a numerical measure computed from a sample (e.g., sample mean or variance). *Sampling distribution* is the probability distribution of a concrete statistic using a random sampling method. For example, the sampling distribution of the mean is the probability distribution of sample-averages. Sampling distributions enable quantitative statistical inference (Maxwell & Delaney, 2004).

Many probability and statistics instructors consider the presentation of theoretically-driven (parametric) inference models, along with the empirically-appealing and tractable resampling and bootstrapping-based inference models, tremendously valuable and important. There is also significant evidence that learners of all levels (formal didactic and informal students) find the dichotomy and synergies between model-based and empirical methods for scientific inference appealing, motivating and practically useful (Windschitl, Thompson, & Braaten, 2008). At the same time, there are differences between these two inference paradigms. One of the methods (parametric inference) is amenable to by-hand calculations, but may have limited applicability. The alternative approach (bootstrap inference) requires powerful computational resources (hardware and resampling algorithms) and may be intractable for by-hand calculations, yet it has a broader scope of applications. The demonstration of the scope, benefits, limitations and practical aspects of both of these inference methods in diverse scientific curricula will give K-16 learners a significant understanding of, and enable them to critically evaluate, the notions of test-sensitivity, model assumptions and computational complexity.

There have been several endeavors to provide interactive resampling-based computational resources to users via the Internet. Some examples include the StatKey/Lock5 webapp (Lock, 2014), Web-Interface for Statistics Education (WISE) applet (WISE, 2014), StatCrunch (West, 2014), RossmanChance (Roy et al., 2014), JMP (Stephens, Carver, & McCormack, 2014), and iNZight (Wild, 2017). And, there are many more randomization inference software tools that are distributed as free or licensed stand-alone applications (Good, 2013; Mills, 2002; Neuhäuser, 2012). Many of these prior developments have limited scope; may be applicable for either demonstrations or data analytics, but not both; require special installation or environment for deployment; may have limited graphical or computational capabilities; are not platform agnostic; or are not available for community support and expansion. Established software tools like R (Aronow & Samii, 2014; Canty, 2002), SPSS (Hayes, 1998) and SAS (Chaudhary & Moulton, 2006) also provide macros, modules, procedures or packages for randomization-based inference. The main draw-back of these alternatives is the need for programming and software expertise to set up and initiate the resampling process, and then to interpret the results of the experiments in command-line or graphical interfaces. In this manuscript, we present a new randomization-based statistical inference infrastructure (http://socr.umich.edu/HTML5/Resampling_Webapp) which is graphical, runs in a web-browser, is platform-agnostic, blends research-driven data analytics and interactive learning, and provides a powerful backend computational library for managing large amounts of simulated or user-provided data.

Although the target audience for this specific SOCR Webapp are students taking Introductory Statistics, or Data Science, courses, it may also be useful for more advanced or graduate service courses. The dichotomy between the experimental and theoretical is indeed well documented (Prodromou, 2012). The overall SOCR framework includes case-studies, analytical methods, visualization tools, and computational services that address the intricate, and realistic, challenges associated with statistical inference based on both univariate and multivariate data. For instance, (1) SOCR Motion Chart Data Dashboard (<http://socr.umich.edu/HTML5>) provide graphical visualization and interrogation of multivariate datasets, (2) the Data Science and Predictive Analytics EBook (<http://DSPA.predictive.space>) includes model-free methods for forecasting, prediction and clustering of extremely high-dimensional datasets, and (3) the SOCR GitHub partition (<https://github.com/SOCR>) includes a number of case-studies and advanced R-code that address realistically the challenges associated with multiple collinearity, latent effects, and confounding.

Random sampling applies stochasticity, or randomness, in the sampling scheme and reflects what is sampled as well as the underlying distribution of that sample. In parametric-based statistical inference (Lindsey, 1996), the random sampling reflects the stochastic nature of selecting observations from the sample space (Glynn & Iglehart, 1989; Hastings, 1970; Pesarin, 2015). In contrast, in randomization-based inference (e.g., bootstrapping) (Efron, 2003; Ferraty, Keilegom, & Vieu, 2010; Maxwell & Delaney, 2004), the random sampling indicates the resampling and stochastic assignment of units to treatments or groups (Koenig, Melie-García, Stein, Strik, & Lehmann, 2008).

II. Motivational Examples

Let us begin by exploring three motivational examples of parametric-based and randomization-based statistical inference. These examples are chosen due to their common use in probability and statistics courses, their direct applications in applied statistical inference, and the fact that we can illustrate their theoretical, empirical and resampling based properties.

Example 1: Binomial Inference.

Many biomedical experiments involve studies of repeated dichotomous processes (Kleinbaum, Kupper, & Morgenstern, 1982). For instance, an experiment comparing groups of similar insects under various concentrations (treatment conditions) may investigate their mortality (in terms of their numbers or different proportions). In such studies of concentration–mortality relationship, insects can be exposed to increasing concentration levels of specific biochemical agents, and investigators may observe the effects in terms of insect survival. For simplicity, let's assume that 50% of the insects are expected to survive an experiment under normal conditions (no treatment) and we perform an experiment using the treatment to observe proportion of survivals. A common type of statistical inference in this situation would be to compare the insect mortality results in the treatment group to an expected binomial model. More specifically, if the treatment sample includes 20 insects and 15 survivors at the end of the experiment, would this difference of 5 survivors more than expected (10, representing 50% of 20) be statistically significant?

The Binomial probability model may be used to theoretically compute the likelihood that the observed difference of 5 survivors is simply due to chance alone. Let the variable X represent the number of surviving insects, and assume that about half of the insects are expected to survive under normal (no treatment) conditions. Then, the probability distribution of X would be $B(n = 20, p = 0.5)$. The mathematical model for this experiment can be presented as flipping a fair coin (the probability of a head turning up is $p(H)=0.5$) 20 times and observing an outcome containing 15 heads, when the expectation is 10 heads. A generic question in this type of situations is “*Is this outcome atypical?*” A more specific question could be “*What is the chance of observing 15 or more heads in this experiment (when the expected number of heads is 10)?*”

Figure 1

The exact probability of observing 15 or more heads is $P(X \geq 15) = 0.020695$ (SOCR, 2014b). This theoretical probability can also be empirically estimated by running 1,000 *Binomial*(20,0.5) simulations and observing the number of outcomes with 15 or more Heads (Distributome, 2014). One such simulation using the Probability Distributome simulator (<http://www.distributome.org/V3/sim/BinomialSimulation.html>) (I. Dinov, Siegrist, K, Pearl, DK, Kalinin, A, Christou, N, 2015), **Figure 1**, generated a data-driven estimate of probability $P(X \geq 15) \approx 0.02$, which is close to the exact theoretical probability $P(X \geq 15) = 0.020695$. Although these empirical estimates may change slightly from one experiment to the next, the law of large numbers guarantees that they will converge to the theoretical probability (as the sample-size increases) (I. Dinov, Christou, & Gould, 2009). In situations where the theoretical probability may not have a (known) closed-form analytical expression, or is computationally intractable, simulation-based inference provides an alternative approach for obtaining useful probability estimates for practical applications.

Example 2: Differences in Proportion.

Human health research provides many powerful applications of computational statistics to identify associations between subject phenotypes, genetic traits, clinical treatments and health outcomes. Studies of heart attacks, Acute Myocardial Infarction (AMI), provide one specific example where mortality rates between the two genders can be compared. This dataset (SOCR, 2013) includes information about all hospital discharges in New York State of heart attack patients who did not undergo heart surgery, in 1993. The sample size is 12,844 patients and the variables included in the dataset are summarized in **Table 1**.

Table 1

Suppose we are interested in comparing mortality rates between females and males. Let p_f = proportion of female patients that die and p_m = proportion of male patients that die. We are interested in testing a null hypothesis $H_o: p_f - p_m = 0$ against an alternative research hypothesis $H_a: p_f - p_m \neq 0$. **Table 2** contains the distribution of the outcomes by gender.

Table 2

The test statistics $Z_0 = \frac{\text{Estimate} - \text{Hypothesized Value}}{SE(\text{Estimate})} = \frac{\hat{p}_f - \hat{p}_m - 0}{\sqrt{\frac{\hat{p}_f(1-\hat{p}_f)}{n_f} + \frac{\hat{p}_m(1-\hat{p}_m)}{n_m}}} \sim N(0,1)$, where $n_f = 767 + 4298 = 5065$

and $n_m = 643 + 7136 = 7779$. Thus, $Z_0 = 11.60525$ and the corresponding p-value is $< 10^{-12}$. This small probability value indicates that the differences in the mortality rates between females and males are not simply due to chance alone; there does appear to be strong gender bias in the rate of deaths from AMI.

One can also use the parametric Statistics Online Computational Resource (SOCR) Chi-Square (χ^2) test to automatically compute the test statistics and p-value, as well as to compute the 99% confidence interval for the difference of proportions:

$$\hat{p}_f - \hat{p}_m \pm z_{0.005} SE(\hat{p}_f - \hat{p}_m) = 0.06877296 \pm 2.576 \times 0.01526543 = [0.02944921 : 0.1080967].$$

If we are unsure about the parametric assumptions (e.g., the sample sizes (n_m and n_f) are large, relative to the (unknown) population proportions (p_m and p_f), i.e., the products $n_m p_m$, $n_f p_f$, $n_m(1 - p_m)$, and $n_f(1 - p_f)$ are relatively large), randomization-based inference may be employed to investigate if the gender effects on mortality rates are significant.

Next, we illustrate the randomization-based statistical inference using the clinical Acute Myocardial Infarction (AMI) cardio-vascular data. The goal will be to identify between-group differences (gender effects on the dichotomous clinical outcome, survival or dying) using the SOCR randomization webapp. As in the previous parametric analysis, we denote p_f = proportion of female patients that die and p_m = proportion of male patients that die, and we are interested in testing a null hypothesis $H_0: p_f - p_m = 0$ against an alternative research hypothesis $H_a: p_f - p_m \neq 0$, this time using the non-parametric resampling-based inference. **Figure 2** shows one instance of the simulation using $K=5,000$ iterations. The resulting resampling results will vary with each resampling experiment. However, the test-statistics and p-value will remain stable, as the number of simulations is large. The almost trivial probability value of the resampling test indicates that observed differences in mortality rates between males and females in this cohort of cardio-vascular patients are not likely to be driven by chance alone, and that gender is a significant factor.

For each iteration of this randomization experiment ($K=5,000$), we generate two random samples ($n_1 = 5,065$ and $n_2 = 7,779$), each corresponding to the 5,065 females and the 7,779 males in the original dataset. These simulations are generated by mixing all cases and randomly extracting two groups of the specified sample-sizes by resampling (with replacement) from the pooled data. Then, for each pair of random samples, we compute the difference of proportion of people that survive or die, the corresponding Z score and the p-value. The sampling distributions of both the test statistics and the corresponding p-values are shown on the left side of **Figure 2** (note that the F-statistics for 2 groups coincide with the t-statistics). The right side of **Figure 2** depicts all of the actual random samples ($K=5,000$) for each of the two groups of sizes $n_1 = 5,065$ and $n_2 = 7,779$.

The most important point of this difference-of-proportions example is the agreement between the parametric (Z -test for proportions) and the webapp-based (resampling simulation) approaches. This consensus, in terms of the final inference on the significance of gender-effects in the mortality of the Acute Myocardial Infarction patients, is not surprising in light of the large sample size.

Figure 2

Example 3: Group comparison inference.

A common application in many scientific studies involves comparing the differences between the distributions of multiple groups or populations. Using sample data from multiple groups, one may compute a set of corresponding group-wise sample statistics (e.g., sample means or medians), which provide the basis for a statistical analysis quantifying the random chance of observing the specific group differences indicated by the sample statistics. As a generalization of the previous example for two or more groups, this example is an extension of the AMI case above. Note that for two or more groups, the webapp defaults to computing the F-statistics for the differences of the group means, as it naturally agrees with the more standard T-statistics used for comparing the means of two independent groups.

Technical implementation details about the implementation and the core features of the randomization and resampling inference webapp are presented in the **Supplementary Materials**.

III. Hands-on Learning Activities

Below we demonstrate two use-cases of the randomization webapp: an *Exploratory* study, based on simulated data, where the user generates sample data and interacts with the webapp in a data-inquiry manner; and an *Explanatory* or *confirmatory* study based on user-specified data, where specific *a priori* hypotheses can be tested. Details about these are also available online (SOCR, 2014j).


1) Exploratory Use-Case – Generating data and using simulations for quantitative statistical inference, **Figure 3**.

- Load the webapp in a modern browser (http://socr.umich.edu/HTML5/Resampling_Webapp)
- Using the Coin-Toss experiment, generate a test dataset by clicking “*Binomial Coin Toss*”
- Choose the parameters – number of coins, probability of Heads, and number of groups (e.g., $k=4$)
- Click “*Generate Dataset*” (you can click this button multiple times, notice how the data samples change)
- Click “*Generate Random Samples*”
- In STEP 2: Generate random samples from selected datasets, enter the number of samples you require, e.g. 10,000
- Click the “1 Sample” button (for one simulation) or the “Generate” button (for larger number of simulations)
- You can inspect all samples (for the k groups) in the right panel of the webapp (use the “Show” button and inspect all the glyphs on the top)
- In **STEP 3: Choose an inference**, select the test statistic you require, e.g. p-value, and click GO
- This will automatically open the “Inference Plot” tab where the randomization distribution of simulated proportion of heads is shown and the initial p_o value is drawn on top to show the relation to the resampling-based distribution.
- You can always modify your prior choices in the “Control” tab.

Figure 3

As the coin characteristics are unchanged between the 4 groups of experiments (fair coin is used and the same number of coin tosses is performed in all 4 groups), we do not expect to see significant between-group differences in the number of head outcomes of the completed study. The non-parametric randomization based inference is in agreement with this intuitive expectation.

2) *Explanatory* Use-Case – Statistical Inference on observed data, **Figure 4**.

- Initiate at the randomization webapp and select the “Use a Excel Sheet” option
- Click the “Reset” button () to remove any previous data from the webapp buffer
- Click on the top-left cell (A1) and copy-paste data from any external spreadsheet. The SOCR Data collection (SOCR, 2014e) provides many examples. For demonstration purposes, we assume we are working with the human Heights/Weights dataset. If you copied the column headers, you may need to use the tool-box to select “Use first row as titles”
- Select a set of, say 20, weight measurements and click “Add as dataset (selection)” (this would represent the first sample). Repeat this selection with another set of 20 Weights (to select a second data sample)

- Click “Proceed”. You should see a summary indicating the sample-sizes of the 2 groups of data you selected
- Clicking “Done” will automatically open the “Control” panel
- In **STEP 2: Generate random samples from selected datasets**, enter the number of samples you require, e.g. 10,000
- Click the “Generate” button
- You can inspect all samples (for the $k=2$ groups) in the right panel of the webapp (use the “Show” button and inspect all the glyphs on the top)
- In **STEP 3: Choose an inference**, select the test statistic you require, e.g. p-value, and click GO
- This will automatically open the “Inference Plot” tab where the randomization distribution (of sampling distribution of the difference of means or p-values, depending on the chosen test statistics) is shown and the (raw) initial mean value is drawn on top to show the relation to the resampling-based distribution. If the raw mean value is towards the extreme of the resampling distribution, then there is sufficient evidence suggesting that the grouping effect is real, and that the samples are likely coming from different distributions (i.e., different processes may have generated the initial samples).

Figure 4

As expected in this case, the randomization-based inference indicates that there are no statistically significant differences between the weights of the two cohorts, which are randomly chosen from the same population of human weights. During each experiment, the results may vary based on the selection of the initial samples.

IV. Discussion

In this paper, we presented several examples motivating randomization and resampling-based statistical inference. In addition, we reported on the development of a new, open infrastructure for data-driven or simulation-based statistical inference. The human interface to this non-parametric inference framework is provided by a webapp (http://socr.umich.edu/HTML5/Resampling_Webapp), which is platform-agnostic, blends research-driven data analytics and interactive learning, and provides a powerful backend computational library for managing large amounts of simulated or user-provided data. We demonstrated the parallels between parametric and distribution-free statistical inference using several examples. The newly developed integrated resampling and simulation framework (including data, web-services, graphical resources and statistical computing libraries) is freely available on the web without access barriers. The SOCR Randomization webapp can be invoked in two alternative ways. By toggling on the “Help” menu at the start of the randomization webapp, the user is guided through the inference process using detailed descriptions of the different components of the webapp, e.g., importing data, initiating the resampling protocol, producing the multiple random samples, inspecting the induced randomization sampling distribution, and making the final inference. More experienced users can suppress the “Help” guidance and directly utilize the computational infrastructure without documentation assistance.

We believe the SOCR Randomization and Resampling webapp infrastructure will be useful for three types of audiences. This framework allows modifications, tailoring and expansions of the JavaScript code to meet the specific needs of the instructors, since the webapp runs on mobile devices and provides simulated or data-driven, non-parametric inference. A limitation of the current implementation of the webapp is the lack of multivariate distribution capability, which may

be developed later by our group or others. As all modern data is high dimensional, a future (community) extension of the webapp capability may include the functionality to bootstrap a multivariate linear model.

Instructors may employ the webapp in and out of the classroom for demonstrating different scientific concepts including sampling, random variation, computational statistical inference and data-driven analytics. Informal learners may also find these resources useful as refreshers, computational calculators or validators for alternative parametric-based inference calculations. Finally, researchers and developers may utilize, expand, modify and embed the open randomization resources (data, source-code, learning activity, services) in other projects without any barriers.

V. Acknowledgments

Many SOCR faculty, students and staff have helped and supported the development, deployment and validation of this randomization and resampling webapp and its wrapper activity. Motivation, ideas and source code from Kyle Siegrist (UAH), Dennis Pearl (OSU) and the Probability Distributome project (www.Distributome.org) were instrumental in these developments. This work was funded in part by NSF grants 1734853, 1636840, 1416953, 0716055 and 1023115, NIH grants P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760, and The Elsie Andresen Fiske Research Fund. The open-source webapp development was partly supported by Google® as part of the Google Summer of Code (GSoC 2012) training program.

References:

- Al-Aziz, J., Christou, N., & Dinov, I. (2010). SOCR Motion Charts: An Efficient, Open-Source, Interactive and Dynamic Applet for Visualizing Longitudinal Multivariate Data. *JSE*, 18(3), 1-29.
- Aronow, P., & Samii, C. (2014). RI: R package for performing randomization-based inference for experiments. Retrieved from <http://cran.r-project.org/web/packages/ri/ri.pdf>
- Barber, J. A., & Thompson, S. G. (2000). Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine*, 19(23), 3219-3236.
- Barker, T. (2013). Data Visualization with D3 *Pro Data Visualization using R and JavaScript* (pp. 65-84): Springer.
- Budgett, S., Pfannkuch, M., Regan, M., & Wild, C. J. (2013). Dynamic visualizations and the randomization test. *Technology Innovations in Statistics Education*, 7(2).
- Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance In Medicine*, 35(2), 261-277.
- Canty, A. J. (2002). Resampling methods in R: the boot package. *R News*, 2(3), 2-7.
- Chaudhary, M. A., & Moulton, L. H. (2006). A SAS macro for constrained randomization of group-randomized designs. *Computer Methods and Programs in Biomedicine*, 83(3), 205-210.
- Che, A., Cui, J., & Dinov, I. (2009). SOCR Analyses – an Instructional Java Web-based Statistical Analysis Toolkit. *JOLT*, 5(1), 1-19.
- Che, A., Cui, J., & Dinov, I. (2009). SOCR Analyses: Implementation and Demonstration of a New Graphical Statistics Educational Toolkit. *JSS*, 30(3), 1-19.
- Christou, N., & Dinov, I. (2011). Confidence Interval Based Parameter Estimation—A New SOCR Applet and Activity. *PLoS ONE*, 6(5): e19178. doi:doi:10.1371/journal.pone.0019178
- Cobb, G. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician*, 69(4), 266-282.
- Dinov, I., Christou, N., & Gould, R. (2009). Law of Large Numbers: the Theory, Applications and Technology-based Education. *Journal of Statistical Education*, 17(1), 1-15.
- Dinov, I., Siegrist, K, Pearl, DK, Kalinin, A, Christou, N. (2015). Probability Distributome: a web computational infrastructure for exploring the properties, interrelations, and applications of probability distributions. *Computational Statistics*, 594, 1-19. doi:10.1007/s00180-015-0594-6
- Distributome. (2014). Distributome Binomial Experiment. Retrieved from <http://www.distributome.org/V3/exp/BinomialExperiment.html>
- Dugard, P., File, P., & Todman, J. (2012). *Single-case and small-n experimental designs: A practical guide to randomization tests*: Routledge.
- Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (Vol. 191): CRC Press.
- Efron, B. (2003). Second thoughts on the bootstrap. *Statistical Science*, 18(2), 135-140.
- Ferraty, F., Keilegom, I., & Vieu, P. (2010). On the Validity of the Bootstrap in Non-Parametric Functional Regression. *Scandinavian Journal of Statistics*, 37(2), 286-306.
- Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference*: Springer.
- Glynn, P. W., & Iglehart, D. L. (1989). Importance sampling for stochastic simulations. *Management Science*, 35(11), 1367-1392.
- Good, P. I. (2013). *Introduction to Statistics Through Resampling Methods and R*: Wiley. com.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97-109.

- Hayes, A. F. (1998). SPSS procedures for approximate randomization tests. *Behavior Research Methods, Instruments, & Computers*, 30(3), 536-543.
- Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6), 497-526.
- Ho, A. J., Hua, X., Lee, S., Leow, A. D., Yanovsky, I., Gutman, B., Thompson, P. M. (2010). Comparing 3 T and 1.5 T MRI for tracking Alzheimer's disease progression with tensor-based morphometry. *Human Brain Mapping*, 31(4), 499-514. doi:10.1002/hbm.20882
- Ho, D. E., & Imai, K. (2006). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *Journal of the American statistical association*, 101(475), 888-900.
- Holcomb, J., Chance, B., Rossman, A., Tietjen, E., & Cobb, G. (2010). Introducing concepts of statistical inference via randomization tests. *Data and context in statistics education: Towards an evidence-based society (ICOTS8)*, Voorburg, The Netherlands.
- Kilburn Jr, H., Schoen, L., & Wang, T. (2012). Acute Myocardial Infarction in New York State: 1996–2008. *Journal of community health*, 37(2), 473-479.
- Kintner, E. K., & Sikorskii, A. (2009). Randomized clinical trial of a school-based academic and counseling program for older school-age students. *Nursing research*, 58(5), 321.
- Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: principles and quantitative methods*: Wiley.
- Koenig, T., Melie-García, L., Stein, M., Strik, W., & Lehmann, C. (2008). Establishing correlations of scalp field maps with other experimental variables using covariance analysis and resampling methods. *Clinical Neurophysiology*, 119(6), 1262-1270.
- Kovbasyuk, O., & Blessinger, P. (2013). The Future Of Meaning-Centered Education *Meaning-Centered Education: International Perspectives and Explorations in Higher Education*, 186.
- Lincoln, D. E., Fajer, E. D., & Johnson, R. H. (1993). Plant-insect herbivore interactions in elevated CO₂ environments. *Trends in Ecology & Evolution*, 8(2), 64-68.
- Lindsey, J. K. (1996). *Parametric statistical inference*: Clarendon Press Oxford.
- Lo, J. T. K., Wohlstadter, E., & Mesbah, A. (2013). *Imagen: Runtime migration of browser sessions for JavaScript web applications*. Paper presented at the Proceedings of the 22nd international conference on World Wide Web.
- Lock. (2014). Lock5 Randomization Webapp. Retrieved from <http://goo.gl/wCNA2z>
- Lohr, S. L. (2009). *Sampling: design and analysis*: Thomson.
- Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, 52(2), 127-132.
- Macrae, C. (2013). *Learning from jQuery*: O'Reilly Media, Inc.
- Manly, B. F. (2007). *Randomization, bootstrap and Monte Carlo methods in biology*: Chapman & Hall.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data*: Psychology Press.
- McGinnis, J. R., Hestness, E., Riedinger, K., Katz, P., Marbach-Ad, G., & Dai, A. (2012). Informal science education in formal science teacher preparation *Second International Handbook of Science Education* (pp. 1097-1108): Springer.
- Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, 10(1), 1-20.
- Myers, R. H. (1990). *Classical and modern regression with applications* (Vol. 2): Duxbury Press Belmont, CA.
- Neuhäuser, M. (2012). Nonparametric statistical tests: A computational approach. *AMC*, 10, 12.
- NodeJS. (2014). NodeJS Project GitHub Source Code. Retrieved from <https://github.com/headjs>
- Pesarin, F. (2015). Some elementary theory of permutation tests. *Communications in Statistics-Theory and Methods*, 44(22), 4880-4892.

- Prodrômou, T. (2012). Connecting experimental probability and theoretical probability. *ZDM*, 44(7), 855-868.
- Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, 34(4), 441-456.
- Roy, S., Rossman, A., Chance, B., Cobb, G., VanderStoep, J., Tintle, N., & Swanson, T. (2014). *Using simulation/randomization to introduce p-value in week 1*. Paper presented at the Proceedings of the 9th International Conference on Teaching Statistics.
- Russell, T., & Aydeniz, M. (2012). Traversing the Divide Between High School Science Students and Sophisticated Nature of Science Understandings: A Multi-pronged Approach. *Journal of Science Education and Technology*, 1-19.
- Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: a GLM approach*. London, England: Sage.
- Scheffe, H. (1999). *The analysis of variance* (Vol. 72): John Wiley & Sons.
- Sheikh, K., & Bullock, C. (2001). Urban-Rural Differences in the Quality of Care for Medicare Patients With Acute Myocardial Infarction. *Arch Intern Med*, 161(5), 737-743. doi:oi:10.1001/archinte.161.5.737
- Singh, A. (2012). Ajax Complexity. *International Journal of Computer & organization Trends (IJCOT)*.
- Smith, J. (2011). The Best And Worst Jobs For 2011. Retrieved from <http://www.forbes.com/2011/01/07/best-worst-jobs-2011-leadership-careers-employment.html>
- SOCR. (2011). SOCR US 2011 Job Ranking Data. Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_2011_US_JobsRanking
- SOCR. (2013). AMI NY 1993 Heart Attacks Data. Retrieved from <http://goo.gl/ga1CFx>
- SOCR. (2014a). SOCR Analysis of Variance (ANOVA) Java Applet.
- SOCR. (2014b). SOCR Binomial Distribution Calculator. Retrieved from http://socr.ucla.edu/htmls/dist/Binomial_Distribution.html
- SOCR. (2014c). SOCR Box and Whisker Chart Applet. Retrieved from http://socr.ucla.edu/htmls/chart/BoxAndWhiskersChartDemo3_Chart.html
- SOCR. (2014d). SOCR Chi-Square Analysis Applet. Retrieved from http://www.socr.ucla.edu/htmls/ana/ChiSquareCT_Analysis.html
- SOCR. (2014e). SOCR Datasets. Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data
- SOCR. (2014f). SOCR Human Heights and Weights Dataset. Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights
- SOCR. (2014g). SOCR Kruscal-Wallis Analysis Applet. Retrieved from http://www.socr.ucla.edu/htmls/ana/KruskalWallis_Analysis.html
- SOCR. (2014h). SOCR Randomization and Resampling Inference Framework: Technical Documentation. Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_ResamplingSimulation_Docs
- SOCR. (2014i). SOCR Randomization and Resampling Webapp GitHub Source Code. Retrieved from <https://github.com/SOCRedu/Resampling-Randomization-WebApp>
- SOCR. (2014j). SOCR Resampling and Simulation Based Inference Activity. Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_ResamplingSimulation_Activity
- Stephens, M., Carver, R., & McCormack, D. (2014). *From Data to Decision-Making: Using Simulation and Resampling Methods to Teach Inferential Concepts*. Paper presented at the Proceedings of the 9th International Conference on Teaching Statistics.
- Sturges, H. A. (1926). The choice of a class interval. *Journal of the American statistical association*, 21(153), 65-66.
- Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2015). Combating anti-statistical thinking using simulation-based methods throughout the undergraduate curriculum. *The American Statistician*, 69(4), 362-370.
- West. (2014). StatCrunch. Retrieved from <http://goo.gl/l1BAp5>
- Wild, C. (2017). iNZight - a platform for quick exploration of data and easily understanding statistical ideas

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science education*, 92(5), 941-967.

WISE. (2014). Web-Interface for Statistics Education Randomization Java Applet. Retrieved from <http://goo.gl/w9kbpE>

WorldBank. (2013). WorldBank Dataset API. Retrieved from <http://data.worldbank.org/developers/api-overview>

Tables and Figures

Table 1: A fragment of the New York State heart attacks data (missing values are denoted by "."). Diagnosis Related Group coding 121 (AMIs with cardiovascular complications who did not die), 122 (AMIs without cardiovascular complications who did not die), and 123 (AMIs where the patient died).

Patient	Diagnosis	Gender	Diagnosis Related Group	Died (0=yes)	Charges \$	Length of Hospital Stay	Age
1	41041	F	122	0	4752	10	79
2	41041	F	122	0	3941	6	34
3	41091	F	122	0	3657	5	76
...
12844	41091	M	123	1	.	1	81

Table 2: Patient distribution – outcomes by gender.

Summary		Died		Total	Sample estimates of proportions that died
		0	1		
SEX	F	4298	767	5065	$\hat{p}_f = \frac{767}{767 + 4298} = 0.1514314$
	M	7136	643	7779	$\hat{p}_m = \frac{643}{643 + 7136} = 0.08265844$
Total		11434	1410	12844	

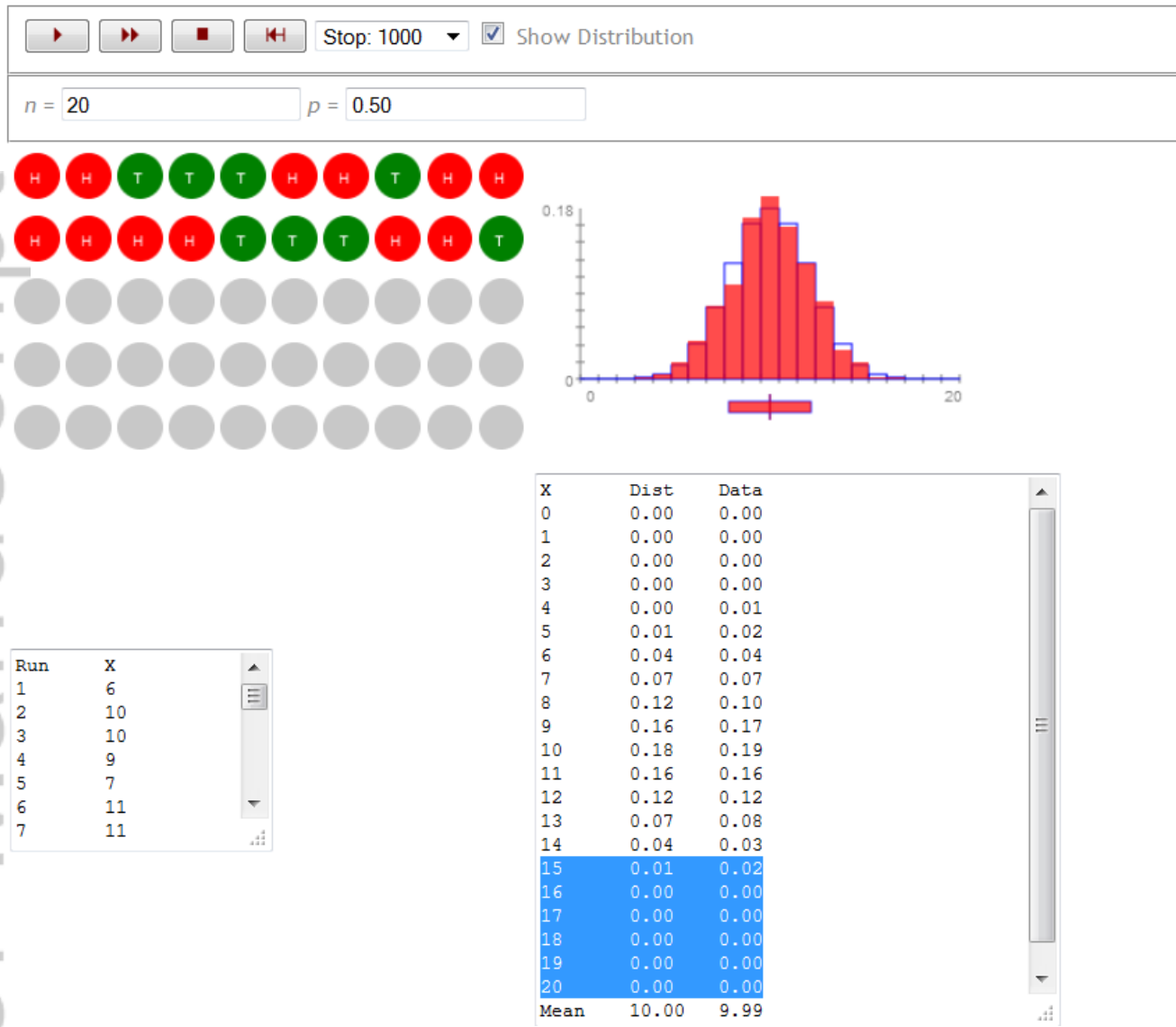


Figure 1: Binomial simulation for estimating the probability that a 20-trial dichotomous experiment, with probability of success equal to that of failure, would generate 15 or more success outcomes, $P(X \geq 15) \approx 0.02$.

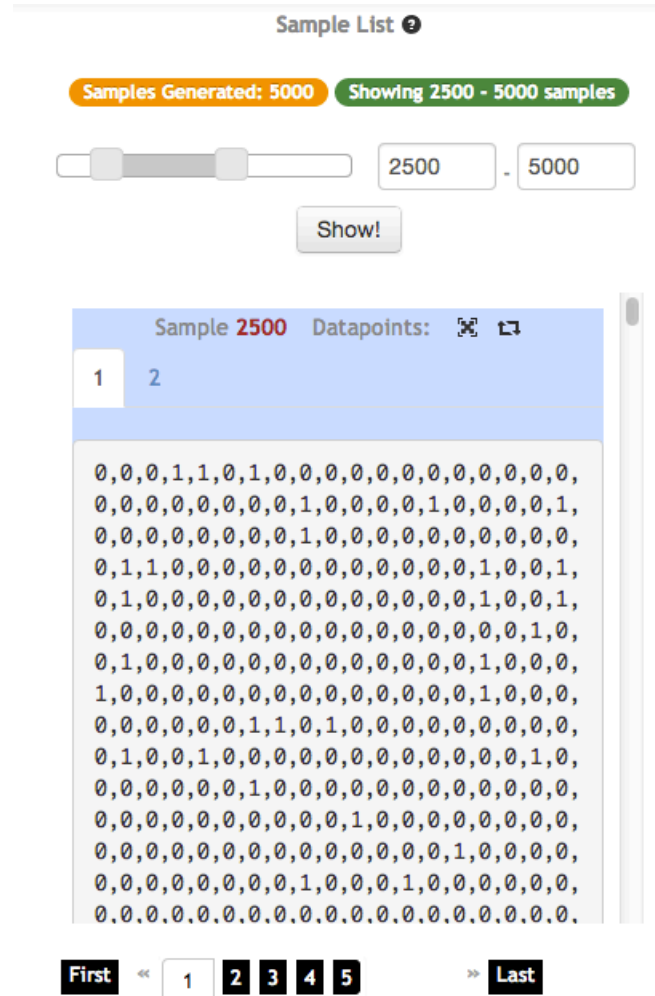
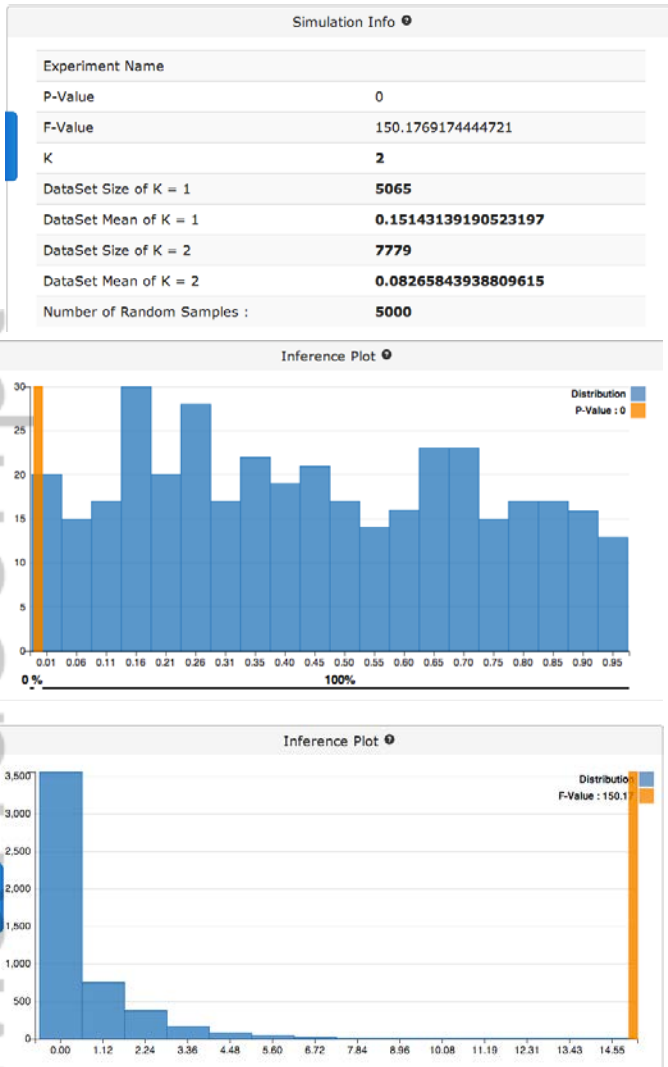


Figure 2: Resampling-based inference results based on K=5,000 simulations. The p-value of the randomization test is approximately equal to zero ($F_{2,K} = 150.18, p \approx 0$), which indicates that there are significant differences between the cardio-vascular mortality rates for males and females in this population.

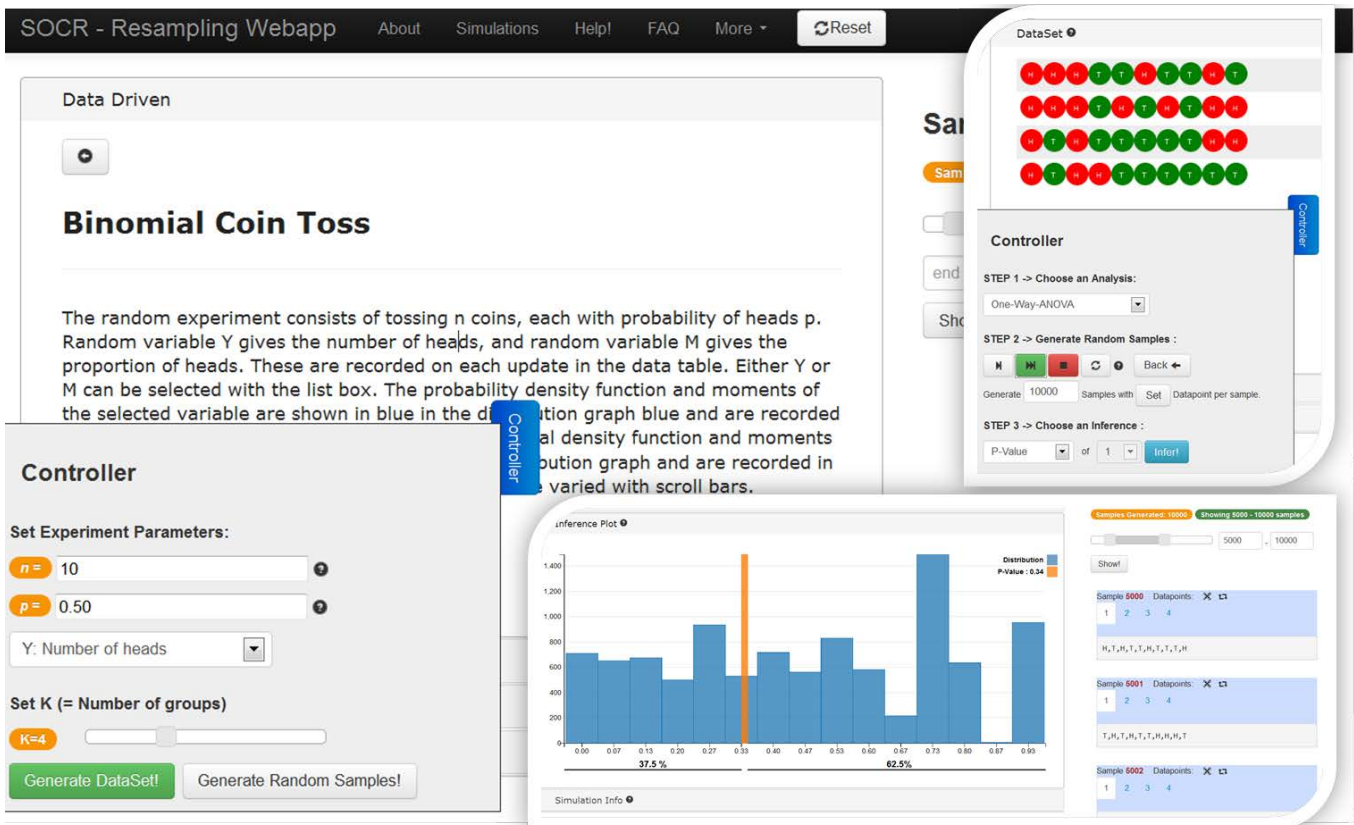


Figure 3: Experiment 1 (exploratory use-case): generating data, performing simulations and completing statistical inference.

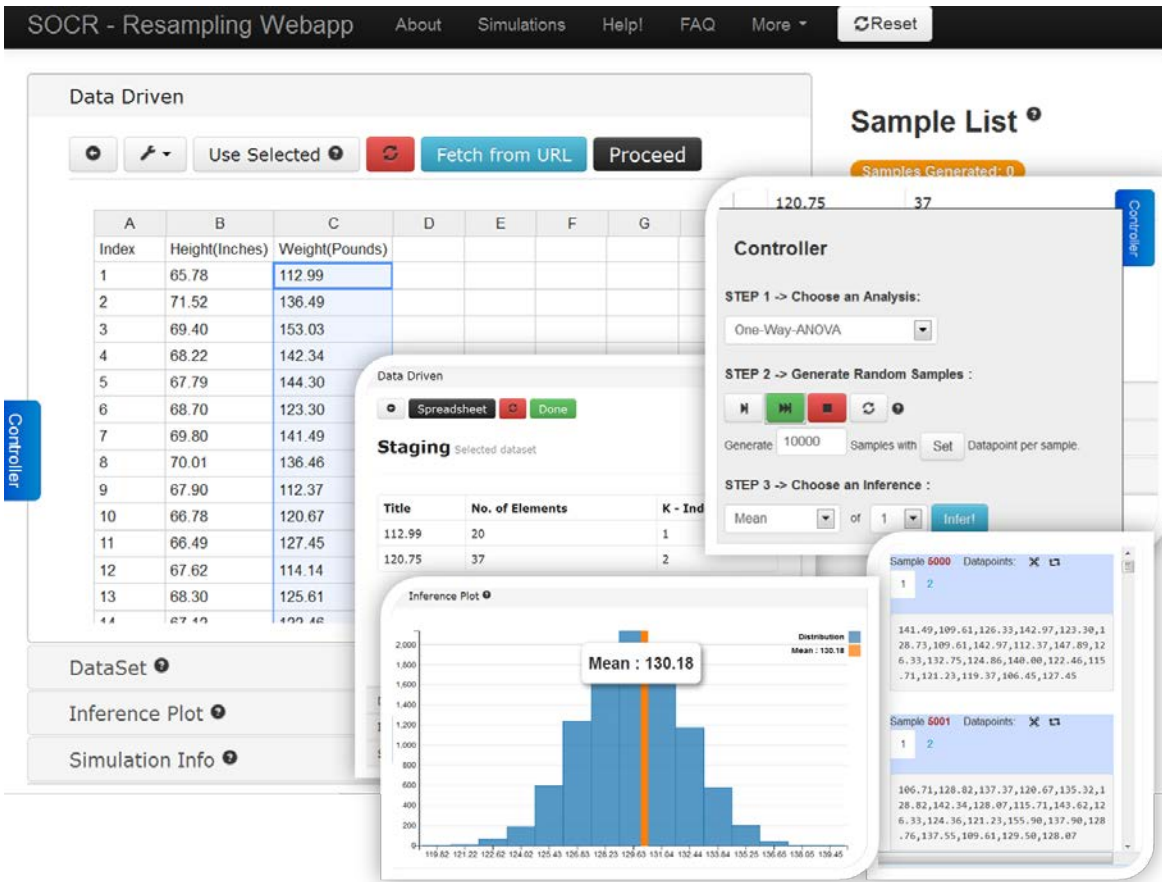


Figure 4: Experiment 2 (explanatory use-case): statistical inference on observed data. This case-study is based on the SOCR human weight and height dataset (SOCR, 2014f) . Once the data is copy-pasted into the webapp data table, we selected 2 random groups of weight measures ($n_1 = 20$ and $n_2 = 37$). However, these settings could be changed depending on the need of the data-driven study. The resampling-based inference indicates that the 2 groups are not different in terms of their mean weights (see orange bar on insert image, which indicates the differences of the mean weights in the original samples, relative to the resampling distribution of differences of randomized group mean weights, blue histogram plot).