FIRSTkit: First Impressions R-based Statistics Toolkit

Israel A. Almodóvar-Rivera [1‡], Ranjan Maitra [2‡],

1 Department of Mathematical Sciences, University of Puerto Rico at Mayaguez Campus, Mayaguez, PR, USA
2 Department of Statistics, Iowa State University, Ames, IA, USA

‡These authors also contributed equally to this work.
¤Current Address: Department of Mathematical Sciences, University of Puerto Rico at Mayaguez Campus, Mayaguez, PR, USA
*israel.almodovar@upr.edu

Abstract

Something will be added here...

Introduction

Modularity is a pervasive concept in computer science, extending from the design of systems (Parnas 1972), to the design of software (Szyperski 1996). Modularity offers several advantages to both a developer and a user. In particular, functionality can be dynamically loaded and unloaded depending on the particular use case. Open source modular software precipitates the possibility of extensions contributed by a wide array of programmers, which can allow the software to morph into areas that weren't anticipated early in development. In the statistics realm, `R` (`R` Core Team 2014) is a prime example of the virtues of modular programming. As of this writing, The Comprehensive `R` Archive Network (CRAN) contains over 9000 source packages which can be installed and dynamically loaded in a particular session as needed.

Modern web technologies have enabled a new generation of software packages that reside solely on the web, which eliminates the issue of local installation and helps abstract away some of the more challenging programming aspects of working directly with `R` . Upon the release of RStudio's Shiny (RStudio and Inc. 2014) it became easier for an `R` -based analysis to be converted to an interactive web application. Several recent software packages have built upon Shiny to provide a web-based system based on R. One such package is iNZight Lite (Wild 2015) which attempts to expose students to data analysis without requiring programming knowledge. Like most web-based systems, this does not include reproducible `R` code, which limits its usefulness in a scientific or academic setting. Another package is called Radiant (Nijs 2016), which is a web-based application with the aim of furthering business education and financial analysis. While the application is modular and extensible, it does require installation and hosting and is inundated with more features than necessary for an introductory student. Partial fulfillment of requirements is noted in the table, as well as a measure of the complexity of functionality offered by default. For example, `R` does have an associated Graphical User Interface (GUI), however this interface is very limited, thus only partially fulfilling the behavior of a GUI.

Unlike other web-based application, `FIRSTkit` is done with the purpose to be used as a teaching component.

As a statistician in a School of Public Health, I saw first hand students struggling with `R` , either by the difficulty or just they weren't interested. The goal of `FIRSTkit` is to allow first time `R` users, or users who have low interest in learning coding in perform basic statistics.

As a web-based application, this tool is immediately more familiar to students than a desktop application. The need for dealing with software licenses, installation configuration, and supported platforms has been eliminated. This allows students to spend more time working with the data and learning statistics than having to struggle to get the software running.

Unlike paid softwares, `FIRSTkit` requires no software licenses or manual installation. `FIRSTkit` 'functionality is focused on introductory statistics students, and nothing more (no extraneous functionality that must be navigated around to get to the content they need). `FIRSTkit` provides students an opportunity to see the underlying functionality of the buttons, text boxes, and other UI elements, in an attempt to foster an interest in coding.

`FIRSTkit` capabilities

The widespread adoption of `R` as a tool for statistical analysis has undoubtedly been an important development for the scientific community. However, using `R` in most cases still requires a basic knowledge of programming concepts which may pose a steep learning curve for the introductory statistics student (Tan, Ting, and Ling 2009). The website is organized around the idea in how introductory statistics courses are carry. `FIRSTkit` capabilities are descriptive statistics, inference and regression. In terms of the descriptive statistics, the user can obtain location summaries as well dispersion.

0.1 Descriptive Statistics

Describing (or summarizing) a data in a clear in a concise way is one of the first things we usually taught in a introductory statistics courses. Many methods are available for summarizing data in both numeric and graphical form. `FIRSTkit` allow to obtain summary regarding location and dispersion. In terms of location these are sample mean, trimmed mean, median, and geometric mean. Given a sample of size n, consider independent random variables $X_1, X_2, \ldots, X_n$ that we wish to summarized.

0.1.1 Location Measurements

1. **Arithmetic mean**: The sample mean from a group of observations is an estimate of the population mean μ. The sample mean is defined to be,

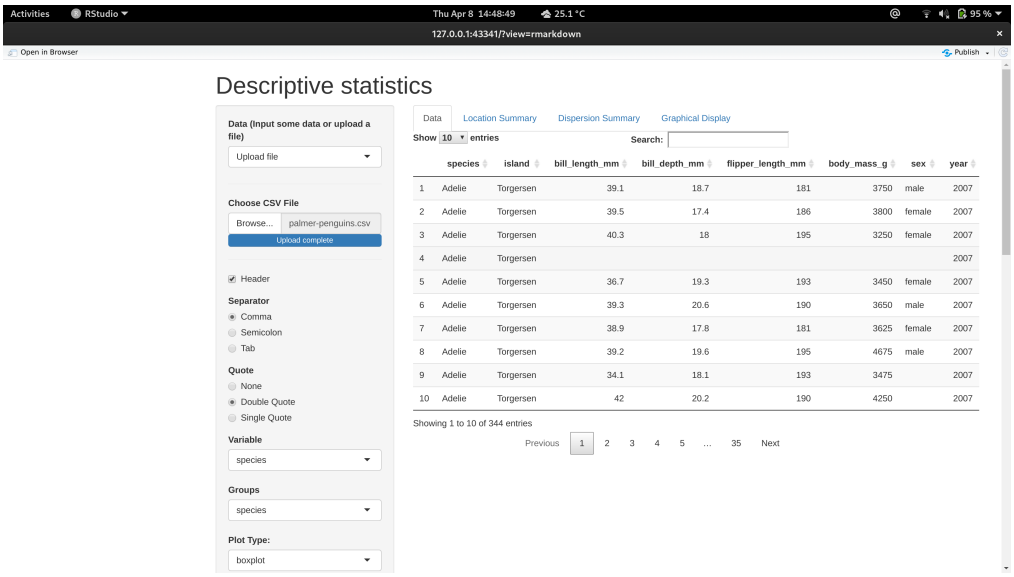
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. Trimmed mean: This mean is computed after discarding given parts of a probability distribution or sample at the high and low end, and typically discarding an equal amount of both. This number of points to be discarded is usually given as a percentage of the total number of points, but may also be given as a fixed number of points.

   (a) First find n = number of observations.

   (b) Reorder them as "order statistic"' X_i from the smallest to the largest.

   (c) Find lower case $p = P/100$ = proportion trimmed.



(d) Compute np. If np is an integer use $k = np$ and trim k observations at both ends. R = remaining observations $= n - 2k$. The trimmed mean is defined as,

$$\bar{x}_k = \frac{X_{k+1} + X_{k+2} + \cdots + X_{n-k}}{R}$$

3. Median: Middle value separating the greater and lesser halves of a data set ,

$$M = X_{(0.5 \times n)}$$

4. Geometric Mean: The geometric mean of a non-empty data set of (positive) numbers is always at most their arithmetic mean. Equality is only obtained when all numbers in the data set are equal; otherwise, the geometric mean is smaller.

$$g = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

0.2 Dispersion Measurements

Given a sample of size n, consider independent random variables $X_1, X_2, \ldots, X_n$, each corresponding to one randomly selected observation. Each of these variables has the distribution of the population, with mean μ and standard deviation σ.

1. Sample standard deviation: is a measure of the amount of variation or dispersion of a set of values.

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

2. Sample Variance: is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of numbers is spread out from their average value.

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

3. Interquartile Range: difference between 75th and 25th percentiles, or between upper and lower quartiles,

$$IQR = X_{(0.75n)} - X_{0(0.25n)}$$

4. Median absolute deviation (MAD): Compute the median absolute deviation, i.e., the (lo/hi) median of the absolute deviations from the median, and (by default) adjust by a factor for asymptotically normal consistency.

$$MAD = \text{median}_i(|X_i - \text{median}(X_i)|)$$

5. Range: the difference between minimum and maximum of all the given arguments.

$$R = \max X_i - \min X_i$$

If `na.rm` is `FALSE`, `NA` and `NaN` values in any of the arguments will cause `NA` values to be returned, otherwise `NA` values are ignored.

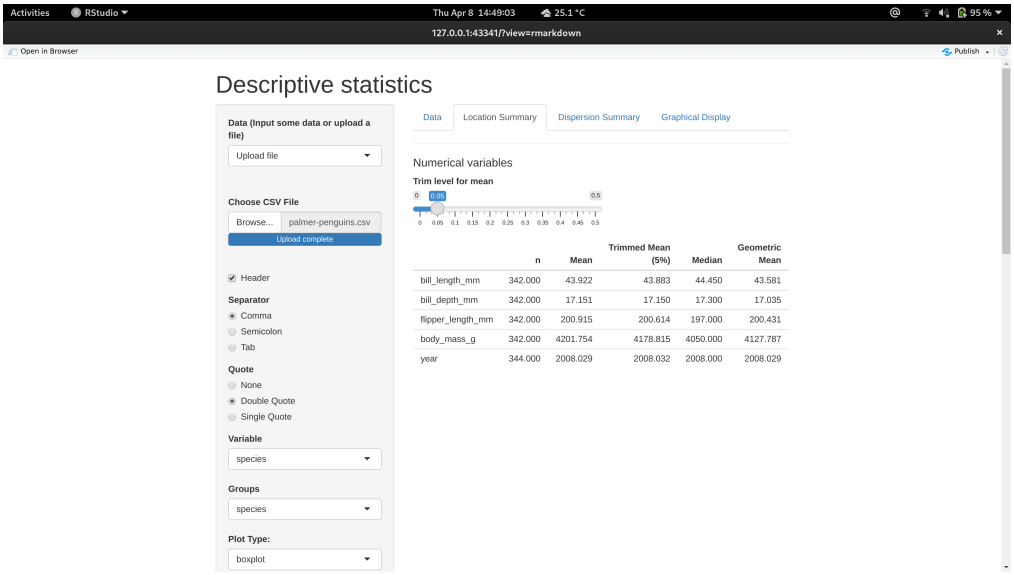


Fig 1. Descriptive Statistics Web-page

0.3 Inference

A chi-square test (Snedecor and Cochran, 1983) can be used to test if the variance of a population is equal to a specified value. This test can be either a two-sided test or a one-sided test. The two-sided version tests against the alternative that the true variance is either less than or greater than the specified value. The one-sided version only tests in one direction. The choice of a two-sided or one-sided test is determined by the problem. For example, if we are testing a new process, we may only be concerned if its variability is greater than the variability of the current process.

0.4 Two Sample t-test

The two-sample t-test (Snedecor and Cochran, 1989) is used to determine if two population means are equal. A common application is to test if a new process or treatment is superior to a current process or treatment.

There are several variations on this test.

- The data may either be paired or not paired. By paired, we mean that there is a one-to-one correspondence between the values in the two samples. That is, if $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are the two samples, then X_i corresponds to Y_i. For paired samples, the difference $X_i - Y_i$ is usually calculated. For unpaired samples, the sample sizes for the two samples may or may not be equal. The formulas for paired data are somewhat simpler than the formulas for unpaired data.

- The variances of the two samples may be assumed to be equal or unequal. Equal variances yields somewhat simpler formulas, although with computers this is no longer a significant issue.

- In some applications, you may want to adopt a new process or treatment only if it exceeds the current treatment by some threshold. In this case, we can state the null hypothesis in the form that the difference between the two populations means is equal to some constant $\mu_1 - \mu_2 = \delta_0$ where the constant is the desired threshold.

0.5 Variance

$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ versus $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

0.6 Proportion

$H_0 : p_1 - p_2 = p_0$ versus $H_1 : p_1 - p_2 \neq p_0$

0.7 Dependent t-test for paired samples

This test is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired". This is an example of a paired difference test. The t statistic is calculated as

$$t = \frac{\bar{X} - \mu_0}{s_D/\sqrt{n}}$$

where $\bar{X}$ are the average and s_D standard deviation of the differences between all pairs. The pairs are e.g. either one person's pre-test and post-test scores or between-pairs of persons matched into meaningful groups (for instance drawn from the same family or age group: see table). The constant µ0 is zero if we want to test whether the average of the difference is significantly different. The degree of freedom used is $n - 1$, where n represents the number of pairs.

0.8 Regression

This Shiny App performs a simple and multiple linear regression. The user can upload (or input) the dataset of interest. User will obtain estimates of the parameters in the model. Also confidence intervals (CI), the default is 95% CI, as well p-value. We can check for the assumptions of normality based on the quantile-quantile plot, assumption of constant variance and linearity. Obtain a fitted model of the simple linear regression:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and $\varepsilon_i \sim N(0, \sigma^2)$.

Conclusions

In this paper, we have outlined a framework for designing a web-based, modular, extensible system which reproduces user actions into `R` code. We believe that the development strategies we've outlined can and should be applied to other software systems, as each of these characteristics aids in the ease-of-use and functionality of the overall product. Although we present them in the context of an introductory statistics application, these ideas are generalizable and we hope that they will gain traction in many other modern software systems.

Since the code is available and these are `.Rmd` files. They can be used as a standalone application.

Acknowledgments

The authors are grateful for the software