iNZight: a gui for learning statistics

Tom Elliott and Chris Wild and Daniel Barnett and Andrew Sporle

October 29, 2020

1 Introduction

- Scope out the need for iNZight

- R can be daunting for beginners/students (who may never have used or even see code before)

- Excel (Microsoft Corporation, 2018), SPSS/etc are rather complex with a not-insignificant learning curve to be able to produce basic explorative plots and summary statistics

    - Do not provide any kind of a 'pathway' to learning R for data science purposes, either

- Other R-based GUIs:

    - Rcommander (Fox, 2005)
    - Jamovi (The jamovi project, 2020)
    - ???

- Other tools targeting students:

    - NZGrapher (Wills, 2020) uses PHP

- No obvious—simple—point-and-click interfaces for simple data analysis/visualisation that also provide a pathway to more complex, code-driven analyses

2 A history of iNZight

- Originally a simple implementation experimenting with R (R Core Team, 2020) and 'gWidgets' (now superceeded by 'gWidgets2', (Verzani, 2019)) for making graphs which react to the type of data (i.e., the user doesn't have to choose the graph type)

- Uses GTK (The GTK+ Team, 2020) to produce graphical interface, accessed via the 'RGtk2' R package (Lawrence and Temple Lang, 2010)

- The software uses the variable types (numeric or categorical) to determine the type of graph or summary produced

- Picked up by [...?] and rolled out for use in NCEA Level 3 statistics in New Zealand (final year of high school)

- Redesigned in 2014 with gWidgets2 and reference classes (one of R's Object Oriented Programming approaches)

- Additionally uses a suite of complimentary R packages to separate form (the UI) from function (data processing, graphics, etc)

- Additionally modules for time series, model fitting, etc, and more recently added an add-on system

- Most of the work has been student-driven: "By students, for students" (rather than being created by computer scientists)

3 An overview of iNZight's structure

- Many of the features come from design-flexibility

- Separation of UI and data processing

- Relies mainly on 'gWidgets2' and 'gWidgets2RGtk2' (Verzani, 2019, 2020) to interface with the user, and reference classes to structure the code base

- internally, each component of the UI is represented by a single class - this makes it easy to extend and modify iNZight

- data/variable manipulation windows each are a single class, which connects to a helper function in another package (usually inside 'iNZightTools' (Elliott, 2020))

- A key design feature is that the UI asks the users for inputs, which are passed to a single function as arguments - this allows not only separation, but also for other GUIs (for example iNZight Lite) to connect to the same functions

- Additionally, those simple wrapper functions can be accessed by beginners not yet ready for the more complex coding required to do the same things

```r
# A simple window to filter a level of a categorical variable
filterWin <- setRefClass('filterWin',
    fields = list(
        GUI = "ANY",
        data = "data.frame",
        variable = "ANY",
        level = "ANY"
    ),
    methods = list(
        initialize = function(gui) {
            initFields(GUI = gui, data = gui$getActiveData())
            win <- gwindow('Filter Level')
            cont <- gvbox(container = win)
            variable <<- gcombobox(names(data), container = cont)
            level <<- gedit("", container = cont)
            okbtn <- gbutton("Filter")
            addHandlerClicked(okbtn, function(h, ...) filterData())
        },
        filterData = function() {
            var <- svalue(variable)
            level <- svalue(level)
            newdata <- iNZightTools::filterLevels(data, var, level)
            GUI$set_data(newdata)
        }
    )
)
```

Listing 1: A simple reference class window to filter a level of a factor.

- examples:
  - `iNZightTools::smart_read()` imports a dataset based on its extension - user doesn't need to know `read_csv()`, `read_dta()`, etc
  - `iNZightPlots::inplot()` is the main power-house function within iNZight - takes UI inputs and generates a graph based on the variable types (and other selections)
  - both of these functions can be accessed directly from R
  - most also return the 'tidyverse' code so learners can get a taste for the actual code necessary to do stuff (e.g., filtering data, etc)

4   Features of iNZight

- Several main groups of functionality:
  - data manipulation (include specifying survey structure)
  - variable manipulation
  - plot control
  - code history

Table 1: iNZight R package family

Package	Description
iNZight	The main package for the GUI
iNZightPlots	Provides plot function 'inzplot()' along with 'inzsummary()' for descriptive statistics and 'inzinference()' for inference and hypothesis testing.
iNZightTools	A suite of helper functions for data process and variable manipulation.

4.1 Data wrangling

- uses 'tidyverse' methods and workflows to perform some data transformation

- all are calls to wrappers inside the 'iNZightTools' package

- the interface allows users to manipulate arguments to the wrapper function - reactive in some cases (i.e., inputs appear/disappear or change values based on previous inputs)

- in many cases, a preview is displayed to help with what can be complex actions, e.g.

- no complex "data tidying" methods - i.e., iNZight expects data to be in rectangular form (but you can move between long/wide formats)

4.2 Graphics and simple data analyses

- all powered by 'iNZightPlots' functions

  - `inzplot()` for graphics
  - `inzsummary()` for simple summary information
  - `inzinference()` for inferencial information and hypothesis testing

- uses the variable type(s) to choose the graph type. Figure x shows the available graph types

- subsetting by 1 or 2 variables is possible/easy/emphasized

  - includes a slider to cycle or 'step' between levels (can add 'motion' to graphs; gapminder example? - great citation opportunity!!)

- plot features: specific to plot type, users only see what's relevant

- some features (such as colour by) might apply to different plot types (scatter and dot plot, for example) and so selection is retained when switching between these plot types

4

- inferencial markup of plots

  - normal theory error bars/curves
  - or bootstrap alternative
  - "Comparison intervals" are a way of visually assessing differences - only shown on differences *that should be compared* - bar plot example
  - an approximate tukey thing? details needed here ('iNZightMR' package)

- summary information

  - a simple numeric summary *of the current plot* (philosophy: look first)

- inference information

  - again, default inferences *of the current plot*
  - both normal and bootstrap options
  - gives a list of relevant hypothesis tests (cf. other software which requires you to decide on hypothesis test first)

4.3   Saving code history

- many functions (the wrappers) generate code and attach it to their result

```
1  data.filtered <- iNZightTools::filterData(iris, "Species", "
     Setosa")
2  cat(code(data.filtered))
3  # iris %>% dplyr::filter(Species == "Setosa")
4
```

- the iNZight GUI stores the attached code after each action and appends it to the R history 'script'

- provides a record of what was done

- helps students interested in continuing data science/statistics to get familiar with code before writing it themselves (can copy+paste and modify, for example)

4.4   Analysing surveys with iNZight

- survey data:

  - iNZight gets told once, and everything else 'just works'
  - plots use survey methods and alternative plot types (histogram vs dotplot; hex bin plot vs scatter plot)

- – summary/inference all use survey methods to obtain summaries/inferences/hypothesis tests

- supports strata/cluster based survey designs, replicate weights, and post-stratification

4.5   Other modules

- A suite of other (fixed) modules for analying/exploring special types of data

  - – time series
  - – multiple response
  - – maps

- or for doing specific tasks

  - – model fitting

- each is its own "reference class" object connecting to one (or more) functions in another package

  - – time series -> 'iNZightTS'
  - – multiple response -> 'iNZightMR'
  - – maps -> 'iNZightMaps'
  - – model fitting -> 'iNZightRegression'

- these are mostly wrappers to other functions, or modified versions -> made to work as 'standalone' packages (i.e., without iNZight)

- not fully implemented, but making progress towards modules also code-writing

- also a newer add-on system allowing anyone to write new modules for iNZight (added manually or through our github-hosted repository)

- easier/faster to maintain/update

- useful for e.g., teachers of a specific course

5   Distributing and running iNZight

- effectively just a collection of R packages (our own, plus many dependencies)

- some on CRAN, others are self-hosted (so windows binaries *are* available)

- most users are new to statistics/data science, most don't know much/anything about R/computing/programming -> we ship an *all-in-one* installer for Windows
  - single directory containing R, iNZight library
  - plus shortcuts to launch R and auto-start iNZight (using `.Rprofile`)
- fully automated build:
  - single repository and github actions
  - builds source/binary files and uploads to our repository
  - builds installer and uploads to website (uses NSIS)
- CRAN release is manual (obviously)

5.1 A note for macOS and Linux users

- GTK support for mac is dead - so we cannot support macOS any longer
- Linux requires installation of some platform-specific packages (gtk, xorg, etc etc)
- just like installing any other R package
- but we do include a build tool to help install and set-up run/update scripts to make launching easier

5.2 Manual installation

- install the packages
- run

```
1  library(iNZight)
2  iNZight()
3  # optionally pass a dataset directly:
4  iNZight(iris)
5
```

6 Discussion and Future Work

- iNZight is a simple-to-use tool for exploring/visualising/analysing data
- built for beginners with capabilities to progress to R-driven coding
- easy-to-learn makes it ideal for organisations that need to do specific jobs rarely (i.e., every few months) - more complicated tools are easy to forget
- flexibility of design makes it easy to add/change functionality as demand changes/grows (e.g., survey design)

6.1 Future Work

- better survey handling

- filling the 'code writing' gaps

- connecting to databases (and potentially processing as much as possible within the db)

-

References

Elliott, T. (2020). *iNZightTools: Tools for 'iNZight'*. Available from `https://cran.r-project.org/package=iNZightTools`.

Fox, J. (2005). The R Commander: A basic statistics graphical user interface to R. *Journal of Statistical Software*, 14(9):1–42.

Lawrence, M. and Temple Lang, D. (2010). RGtk2: A graphical user interface toolkit for R. *Journal of Statistical Software*, 37(8):1–52.

Microsoft Corporation (2018). *Microsoft Excel*. `https://office.microsoft.com/excel`.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

The GTK+ Team (2020). *GTK*. `https://www.gtk.org/`.

The jamovi project (2020). `https://www.jamovi.org`.

Verzani, J. (2019). *gWidgets2: Rewrite of gWidgets API for Simplified GUI Construction*. R package version 1.0-8.

Verzani, J. (2020). *gWidgets2RGtk2: Implementation of gWidgets2 for the RGtk2 Package*. R package version 1.0-7.1.

Wills, J. (2020). *NZGrapher*. `https://grapher.jake4maths.com`.

Acknowledgements