Homework for Introduction to Statistical Computing with R

taught by Nicholas Reich, UMass-Amherst, Fall 2014

due 19 Nov 2014

Preliminaries

This homework assignment is a little bit longer than the other ones because you have two weeks to do it. Steele and I will be able to provide better feedback and assistance if you ask for help far in advance of the deadline. My recommendation is to try to work on it a little bit every few days over the next two weeks. Don't let your R skills get stale!

Your answers to the following questions should be compiled into a PDF report generated by RMarkdown. Submit your report on Piazza with the filename `[your_lastname]_hw12.R`. All functions written as part of this assignment should be saved in a single file called `data_summary_functions.R` and loaded using `source()`. Additionally, you will need to read in the `data-for-hw.rda` file made available for this assignment: link to file. This is a workspace file that contains several different R objects needed to complete this assignment.

Part 1 (2 pts)

If you were handed a set of measurements (let's assume they are numerical observations on some variable), what is the first plot that you would like to see to help you visualize these data? If you were given the opportunity to summarize the observations with five quantitative metrics, what would you choose?

Part 2 (5 pts)

Write a function that takes a numeric vector as input and (a) checks that the vector is of class "numeric" or "integer", and throws an error if it is not, (b) generates the plot you named in Question 1, and (c) prints out the five metrics in the R console, naming clearly what each of them represents. Name this function `data_summary()`.

Part 3 (5 pts)

Run your function on the three vectors called `vec1`, `vec2`, and `vec3`, and print the output for each call to your `data_sumary()` function.

Part 4 (2 pts)

Run your `data_summary()` function on the `vec4` object, which contains some missing data. After looking at the data, what value(s) do you think are used to designate missing values?

Part 5 (5 pts)

Create a new version of your `data_summary()` function, called `data_summary_missing()`. This function should have the same outputs as the original. Add an argument for this function called `na_value` in which the user specifies values that should be considered missing and turned into NAs. Additionally, create a new argument called `na_rm`, which defaults to FALSE, and which is passed as the `na.rm` argument to any of the functions used to generate summary metrics.

Part 6 (5 pts)

Run your `data_summary_missing()` function on the `vec4`, `vec5`, and `vec6` objects, and print the output for each call to your `data_sumary_missing()` function.

Part 7 (10 pts)

For 1-dimensional data a histogram can serve as a useful summary of the distribution of the data. In 2 dimensions, nonparameteric density estimates tend to be favored over histograms. However, a 2-D histogram can be useful because of its simple interpretation.

Write a function called `bin2d` which constructs a 2-dimensional equally spaced grid surrounding the data points and tabulates the number of data points falling into each grid cell. The function should have the following prototype:

```
bin2d <- function(x, y, nbin) {
    ## body of function
}
```

The arguments x and y are numeric vectors indicating the x- and y-coordinates of the data and nbin is an integer specifying how many cells the grid should have in one of the dimensions. For example if nbin = 4, then bin2d should use a 4×4 equally spaced grid. The limits of the grid should at a minimum cover the range of the data.

The function should not plot or print out anything, but should return an object containing the relevant tabulation information. The specifics of the object returned by bin2d should be determined by you. Please write one paragraph describing the object that is returned by your implementation of bin2d. [1]

Additionally, show the results from running your function on `(vec1, vec2)`, `(vec2, vec3)`, and `(vec1, vec3)`.

Part 8 (2 pts)

Create a user account on github.com. Include your username in your writeup, next to your name at the top of the document.

Extra credit

1. At any time during the course, if you find an error (even as simple as a typo in any of the materials) you may correct it on GitHub and receive extra credit. Specifically, you will need to "fork" the nickreich/statComp2014 repository on GitHub, make the change in your copy of the repo, and then submit a "pull request" to me to incorporate your changes into the master branch of the course material. You will receive 1 point added to your final grade for every line of code edited, and successfully pulled into the master branch.

2. Are the results of your `bin2d` function tidy data? Plot the ouput from your bin2d object using `ggplot2` and `geom_tile()`.

3. Adapt your `bind2d` function to take the `na_values` argument that your `data_summary_missing()` function also had. Demonstrate that it works by running `bin2d` on `vec4` and `vec5`.

[1] Acknowledgements to Roger Peng for this question.