Problem Set 1

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2019

Photo by Howard Jackman

Format

- The submission is:
  - A write-up as a PDF document, which includes any plots, diagrams, tables, pictures, and explanations
    - No code
  - In a **separate** "file", submit all code used to derive the results
    - Tarball, zipped directory, lots of individual files w/ self-explanatory titles, etc.
  - Include any original data files or how the data was accessed
    - If you use a internet scraping tool, note the date when you retrieved the data
    - If you can save the data to a file, do so and submit the data file. There is no need to change the format, e.g. HTML, XML,

Software and Data Handling

- As a precursor to doing computer aided statistics, the first problem set will focus on data handling, parsing text, writing code, and simple presentation
- Exercises will focus on USA college basketball statistics from the 2014 Ken Pomeroy Basketball page at http://kenpom.com/index.php?y=2014
  - The content is largely irrelevant and was chosen due to some *interesting* features
- This will be *potentially* time-consuming
  - Took me 4 hours to do the first time and 2 hours to redo
  - Could be done in ~20 min. if you've already mastered: regular expression package, HTML/XML parser, class declaration, plotting commands, etc…

Assignment Overview

- Conceptually this is a simple assignment
  - No advanced or even difficult statistical methods or analysis
- The goal of the first assignment is to assess how well people can load, analyze, plot data, and write simple explanations of how they achieved the results
  - Essentially a plotting and data throughput exercise
  - But, there are some interesting data features
- Words of advice for the following problem set
  - Don’t be overly reliant on spreadsheets
  - Don’t assume that the input data (or format) is stable between years for exercises 2 and 3
- There are some known (at least by me) ambiguities in the exercises. If you come across what you perceive is an ambiguity, detail it in your write-up.

Exercise 1 (3pts.)

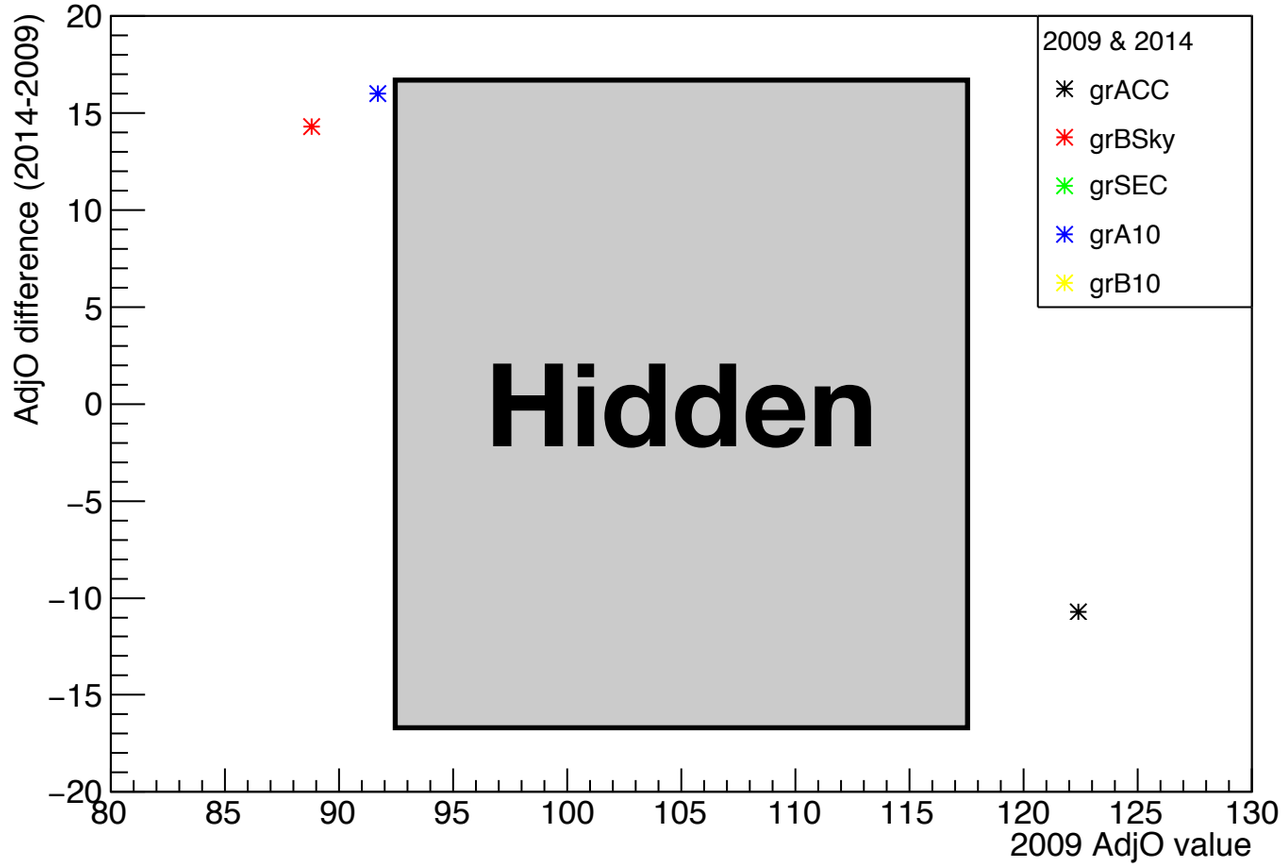
- Take the 2014 Ken Pomeroy data related to NCAA College Basketball analytics from http://kenpom.com/index.php?y=2014
- On a single plot, draw histograms of:
  - The Adjusted Defense "AdjD" for all the teams in the 5 conferences (ACC, SEC, B10, BSky, and A10)
  - Different colors for each conference and add a legend
  - Save as a PDF

Exercise 2 (4pts.)

- Take the 2014 and 2009 Ken Pomeroy data related to NCAA College Basketball analytics
- Calculate the difference in "AdjO" for every team in the 5 conferences from Exercise 1:
  - 2014 minus 2009 as a function of the 2009 AdjO value
  - Plot the data as a graph with a data point for each team entry being the same conference color as for the previous histogram in Exercise 1
- Calculate the difference in "AdjO" for all the teams with data in both 2009 and 2014
  - Median and mean for each of the 5 conferences
  - Median and mean for teams that were not in the 5 conferences

Exercise 2 (example)

- Calculate the difference in "AdjO" for every team in the 5 conferences from Exercise 1:
  - Between 2014 and 2009 as a function of the 2009 AdjO value
  - Plot the data as a graph with a data point for each team entry being the same conference color as for the previous histograms



*Be mindful that this plot is an example and is not guaranteed to be accurate

**Each team should have it's own data point. There are many more points beneath the 'Hidden' box

Exercise 3 (3pts.)

- Take the 2014 and 2009 Ken Pomeroy data related to NCAA College Basketball analytics
- Redo Exercises 1 and 2, while now adding the "BE" conference to the previous list of 5 conferences
  - For those who have written robust code, this should be trivial
  - It is likely to be much harder for those who…
    - Parse some data in by hand
    - Only wrote code that requires the exact data format specific to the team names, conferences, AdjO/AdjD position, etc.

Problem Set Submission

- Due by Wednesday Feb. 13 at 08:30 CET
- In a single submission:
  - Submit the results, plots, numbers, text, etc. in a **single** PDF document
    - The submitted PDF document should not contain any code
  - In a separate file include the code, however terrible, broken, crashing, unpretty, or uncommented in the same submission
  - Unless you parse directly from the internet HTML, also include the data files you actually used. Sometimes files can change, so please supply the one you are actually using.

Exercise 4 (Extra 1pt.)

- One of the most important observations in astronomy was recently made with the coincident observation of gravitational waves in addition to photons across a wide range of wavelengths from a binary neutron star merger
- There is an author list at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2018/data/authors-acknowledgements-v5.pdf
  - How many unique authors are there in that list?
  - If there was one single author list in alphabetical order (instead of being grouped by experimental collaboration), what author is the mid-point
    - Who is at the location (total authors)/2. Potentially there are two authors depending on whether the total number of authors is an odd or even number.