PUI 2015 HW 3. Due W 10/7 at 5:00PM. Automatic scripts will fork your repositories at deadline time, no excuses or delays accepted.

Reading:

Estimating the Reproducibility of psychological sciences
Open Science Collaboration
Science 349, (2015)
DOI: 10.1126/science.aac4716
http://www.sciencemag.org/content/349/6251/aac4716.full.pdf

Assignment 1 : Distributions

Following the ipython notebook
https://github.com/fedhere/PUI2015_fbianco/blob/master/citibikes/citibikes_compare_distributions.ipynb

Test the correlation of the age distribution of Male vs Female riders and of day vs night riders

Use: **age of bikers for male and age of bikers for female gender** for whichever month, or set of months you choose.
test the correlation/independence of the 2 samples using:
Pearson's test (answer: are the 2 samples correlated?)
Spearman's test (answer: are the 2 samples correlated?)
K-S test (answer: are the samples likely to come from the same parent distribution?)
State in words what the tests tell you.

Use: a**ge of bikers during the day and during the night** hours for the same sample .
Test the correlation/independence of the 2 samples using the same tests.
State in words what the tests tell you.

Extra credit: code up your own version of the KS test and compare your result with the one obtained from scipy.stats.ks_2sample

Delivery: check the ipython notebook in the Github directory called CitiBikes you use for this week's assignment. If you work in groups, which is encouraged as usual, state your contribution in the README for the repo.

75% of the grade will be based on the rendered version of the plot, 25% will be awarded if the TA can download and run the notebook (if you include any package that was not in the standard Anaconda distribution state that in your README.md, so that the TAs can install them).

Assignment 2 : **Goodness of fit testing (synthetic data)**

Follow and understand the notebook
https://github.com/fedhere/PUI2015_fbianco/blob/master/HW4/goodness_of_fit.ipynb

Test the difference between a Binomial distribution and a Gaussian distribution as you change the parameters increasing the value of of the Binomial mean (n*p)

Generate a series of binomial samples with increasing n*p (the mean) value. You can choose to change n, p, or both. Test how good an approximation to each sample a Normal distribution is using the KS test (scipy.stats.kstest) the Anderson Darling test (scipy.stats.anderson) and either the K-L divergence of the Chi-square test for goodness of fit (for the latter 2 a bit more coding is required as scipy does not provide functions that accept a sample and the name of a comparison distribution like scipy.stats.kstest, scipy.stats.anderson for these other tests.) Examples in the notebook lead you through the steps described above for the Binomial distribution, but you can choose to change them as you wish.

Plot the measure of similarity you obtain from each test as a function of the parameter values used to generate the sample and explain your result and the visualizations you came up with in a “Caption-like” description of the plots.

Repeat this entire procedure for the Poisson distribution.

Delivery: create a new Github directory called HW4 and check the ipython notebook in it, with a README that states your contribution if you worked in group (which is encouraged).

Assignment 3: Goodness of fit on CitiBike data

Follow and understand the notebook
https://github.com/fedhere/PUI2015_fbianco/blob/master/citibikes/citibikes_goodness_of_fit.ipynb

Test whether a gaussian model for the age distribution of CitiBike drivers is a sensible model, and **if you dare**, try and find a better fit with another distribution! Use 2 tests (from the previous exercise) to do this. Test at least 2 distributions.

Divide your riders sample by seasons: Spring+Summer vs FallWinter. Test how well a normal distribution describes each sample of age distributions. Choose 2 test between: KS, Anderson Darling, Chi2, KL, or any other test for goodness of fit! If you have other ideas let me know and, unless there is a conceptual objection to using the test you are thinking of you can use whatever you want.

Choose a second functional form for the comparison (poisson, gamma……. whatever other distributions). Is it better or worse than the normal distribution fit?

Optional (extra credit): Divide your sample geographically: by Borrow + split Manhattan in an Uptown and a Downtown sample (use your discretion to choose the separation line) and see if you notice any differences in how the age distribution can be modeled.

Delivery: check the ipython notebook in the citibike github repo. usual rules for grading apply.