Graduate Seminar-
A 780
Astrostatistics and Scientific Computing

What you should have on your computer

❖ R

    ❖ Rstudio

❖ Python

    ❖ Ipython

    ❖ Numpy (multidimensional ndarray)

    ❖ scipy

    ❖ astroML (library from Izevic et al. book)

    ❖ scikit-learn

    ❖ pandas (python data analysis library)

❖ Jupyter notebook

❖ Solstice (laptop, ipad)

R and Python

Both.

Each project will have to be completed in both languages.

Project 1

❖ File proportionfile1.dat (two columns of 0s and 1s)

    ❖ First column: test the null hypothesis

        H_0 = the fraction of 1s, f1, is equal to 0.6 (e.g. fraction of spiral galaxies in a galaxy group).

        Alternative hypothesis f1 different from 0.6

    ❖ Test the null hypothesis H_0, f1 of first column=f1 of second column (e.g. comparison of fraction of spiral galaxies in two galaxy groups; or comparison of fraction of spiral galaxies in the inner and outer regions of a galaxy cluster).

    ❖ Set the significance level, calculate the p-value and write the 95 % confidence interval. Write/discuss your conclusion.

❖ Repeat for file proportion2.dat

Project 1

One- and Two-sample mean hypothesis test

One- and Two-sample proportion hypothesis test

Project 1

- ❖ File meanfile1.dat:

  - ❖ First column: test the null hypothesis H_0 = the mean, m, is equal to 0.8. Alternative hypothesis=

    - ❖ m different from 0.8

    - ❖ m larger than 0.8

    - ❖ m smaller than 0.8

  - ❖ Test the null hypothesis H_0,

    H_0 : mean of first column, $m1$=mean of second column, $m2$; alternative hypothesis $m1$ different from $m2$.

  - ❖ Set the significance level, calculate the p-value and write the 95 % confidence interval. Write/discuss your conclusion.

- ❖ Repeat for file meanfile2.dat

Project 1

❖ Each tar file (meanfiles-A.tar, meanfiles-B.tar,…meanfiles-E.tar) includes 100 files (for example meanfiles-A.tar contains meanfilesA1, meanfilesA2, etc.)

❖ For <u>each file in each tar file</u>, test the null hypothesis H_0, mean of first column, $m1$=mean of second column, $m2$; alternative hytpothesis $m1$ different from $m2$

   ❖ Set the significance level, calculate and save the p-value and the 95 % confidence interval of each test.

   ❖ Report, for each tar file, the number of tests resulting in a statistically significant difference.

❖ Illustrate and report your results with plots and tables (create a different section in your report for each tar file). Write/discuss your conclusion.

Project 1

❖ Some useful R functions

  ❖ read.table (to read files)

  ❖ prop.test, t.test

  ❖ For help on any R function help(…): e.g. help(t.test)

❖ Some useful Python functions:

  ❖ scipy.stats.ttest_ind, scipy.stats.ttest_1samp

  ❖ statsmodels.stats.proportion.proportions_ztest

❖ Or write your own functions for the tests.

Project 1-Timeline

❖ Progress report discussion : **September 6**

❖ Final written report and brief presentation (5-10 minutes) of Project 1 results: **September 13**.

❖ Additional questions for Project 1 will be discussed and assigned in class: **September 13**

❖ Answers to additional questions to be included as an Appendix to the Final Report and discussed in class: **September 20**