Scipy: scipy.stats improvements

Organization

Scipy

Personal information

Name: Abraham de Jesus Escalante Avalos
E-mail: aeklant@gmail.com
Github: aeklant
Blog:
GSoC Blog RSS feed:
Telephone: Provided upon request (through more private means)
Timezone: GMT -6
University: The University of Sheffield (England)
Major: Software Systems and Internet Technology
Start date: 28th September 2015
Expected graduation date: July 2016
Degree: MSc
Additional Information:

Personal Bio

I am Abraham Escalante, a Mexican Engineer graduated in 2010 from the Institute of Technology and Higher Education of Monterrey (ITESM, for its acronym in Spanish) with a BS degree in Computer Science. I worked at HP, using Python in the Internal Search department from 2011 to late 2013 but then decided that I needed to be part of something different so I quit my position to join a startup company where I worked developing mobile applications, desktop applications and web e-commerce applications for small to medium sized companies while I followed my dream of studying a master's degree in Europe.

I have now secured my place at **The University of Sheffield's MSc in Software Systems and Internet Technologies** which I am excited to start in September, 2015. I am also proud to say that with the help of the **SciPy** community members I have recently started making contributions to the world of Open Source which is something I would like to continue to do, regardless of the path that my professional career takes in the future.

Project

Abstract

scipy.stats is one of the largest and most heavily used modules in Scipy. With the upcoming release of “Scipy 1.0” it must be ensured that the quality of this module is up to par and while the efforts to improve it have been ongoing, there are still some milestones to be reached in order to accomplish the goal. These milestones include the improvement of documentation, test coverage and code enhancement, most of which is already outlined and described by the community in the form of open issues (1) or proposed enhancements (2).

Deliverables

- All hypothesis tests should get a keyword `alternative` to give the user a choice of hypotheses to use.
- All issues labeled with the StatisticsCleanup milestone must be properly dealt with and closed; this may include documentation, test coverage or deprecation, depending on the issue (1).
- Functions that return multiple parameters must return NamedTuples.
- `scipy.stats.mstats` functions must be made consistent with their `scipy.stats` counterparts.
- `_chk_asarray` and `_chk2_asarray` in `scipy.stats` and `scipy.stats.mstats` must be enhanced to properly handle scalars (3)
- Unit tests for scalar input must be added to for all the functions that make use of `_chk_asarray` and `_chk2_asarray` (3)
- All Python code added or changed as part of the project must be PEP-8 compliant (this includes test code; exceptions may be made in the name of clarity and common sense).

Proposal Timeline

Community bonding period (28th Apr - 24th May)

- Define the scope of all 39 issues listed under the StatisticsCleanup milestone (1).
- Investigate all the hypothesis tests and the details of the `alternative` keyword.
- List all the functions that must return NamedTuples and study their implementations; brush up on NamedTuples (2).
- List all the `scipy.stats.mstats` and define their changes in implementation to make them consistent with their `scipy.stats` counterparts.
- `_chk_asarray` enhancement to handle scalars properly (optional; see issue 4550).
- Implement unit tests to scalar values for all the functions that make use of `_chk_asarray` (optional; see issue 4550).

Week 1 (25th May - 30th May)

- Implement `alternative keyword` addition to all hypothesis tests

- Implement the change to make all functions that return multiple values, return them in a NamedTuple

Week 2 (31st May - 6th Jun)

- Change `scipy.stats.mstats` functions to be consistent with their `scipy.stats` counterparts

Week 3 (7th Jun - 13th Jun)

- Trimmed statistics functions have inconsistent API (issue #2914)
- Statistics Review: fligner (Trac #166) (issue #693)
- Statistics Review: bartlett (Trac #162) (issue #689)
- Statistics Review: ansari (Trac #161) (issue #688)
- Statistics Review: shapiro (Trac #158) (issue #685)

Week 4 (14th Jun - 20th Jun)

- Statistics Review: ppcc_plot (Trac #152) (issue #679)
- Statistics Review: ppcc_max (Trac #151) (issue #678)
- Statistics Review: kstatvar (Trac #149) (issue #676)
- Statistics Review: kstat (Trac #148) (issue #675)
- Statistics Review: bayes_mvs (Trac #147) (issue #674)

Week 5 (21st Jun - 26th Jun)

- Statistics Review: find_repeats (Trac #146) (issue #673)
- Statistics Review: square_of_sums (Trac #138) (issue #665)
- Statistics Review: ss (Trac #136) (issue #663)
- Statistics Review: f_value_multivariate (Trac #135) (issue #662)
- Statistics Review: f_value (Trac #133) (issue #660)

Mid-term Evaluation (26th Jun - 3rd Jul)

- Housekeeping: Make sure all code added or changed so far is PEP8 compliant; catch up with any delayed tasks.

Week 6 (5th Jul - 11th Jul)

- Statistics Review: f_value_wilks_lambda (Trac #123) (issue #650)
- Statistics Review: betai (Trac #115) (issue #642)
- Statistics Review: chisqprob (Trac #114) (issue #641)
- Statistics Review: kruskal (Trac #112) (issue #639)
- Statistics Review: ranksums (Trac #111) (issue #638)

Week 7 (12th Jul - 18th Jul)

- Statistics Review: tiecorrect (Trac #110) (issue #637)
- Statistics Review: mannwhitneyu (Trac #109) (issue #636)
- Statistics Review: linregress (Trac #102) (issue #629)
- Statistics Review: kendalltau (Trac #101) (issue #628)
- Statistics Review: pointbiserialr (Trac #100) (issue #627)

Week 8 (19th Jul - 25th Jul)

- Statistics Review: f_oneway (Trac #96) (issue #623)
- Statistics Review: trim_mean (Trac #93) (issue #620)
- Statistics Review: trim1 (Trac #92) (issue #619)
- Statistics Review: trimboth (Trac #91) (issue #618)

- Statistics Review: threshold (Trac #90) (issue #617)

Week 9 (26th Jul - 1st Aug)

- Statistics Review: signaltonoise (Trac #82) (issue #609)
- Statistics Review: relfreq (Trac #78) (issue #605)
- Statistics Review: cumfreq (Trac #77) (issue #604)
- Statistics Review: histogram2 (Trac #75) (issue #602)
- Statistics Review: describe (Trac #68) (issue #595)

Week 10 (2nd Aug - 8th Aug)

- Statistics Review: variation (Trac #65) (issue #592)
- Statistics Review: moment (Trac #64) (issue #591)
- Statistics Review: masked_var (Trac #58) (issue #585)
- Statistics Review: _chk_asarray (Trac #44) (issue #571)

Weeks 11 and 12 (9th Aug - 21st Aug)

- Housekeeping: All code must be PEP8; make sure all tasks are successfully completed to close the project.
- Pencils down

Code Sample:

https://github.com/scipy/scipy/pull/4619

References and links:

(1) Statistics Cleanup milestone:
https://github.com/scipy/scipy/issues?q=milestone%3AStatisticsCleanup

(2) NamedTuples implementation discussion:
https://github.com/scipy/scipy/issues/4192

(3) Issue 4550 - discussion of `_chk_asarray` proper scalar handling:
https://github.com/scipy/scipy/issues/4550

(4) PEP8 Style Guide for Python Code:
https://www.python.org/dev/peps/pep-0008/