permute: A Python Package for Randomization Inference

Kellie Ottoboni

Department of Statistics, UC Berkeley
Berkeley Institute for Data Science

June 14, 2016

University of California, Berkeley
DEPARTMENT OF STATISTICS

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

Outline

History of randomization inference - fisher - neyman model

R has several packages for randomization inference.

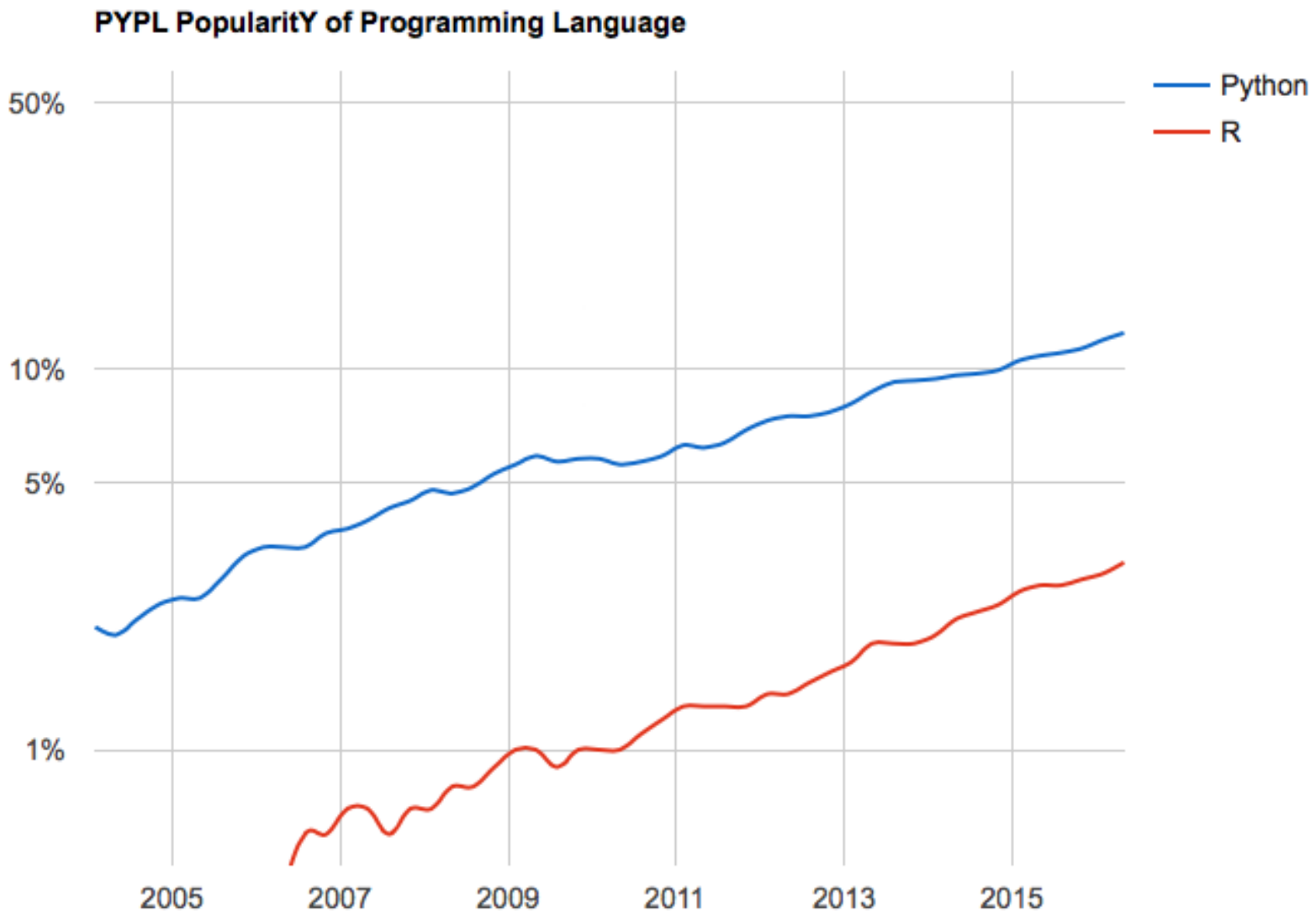
- `ri`
- `RItools`
- `coin`
- `perm`

In Python, statistics packages are limited.

- `numpy.random`
- `scipy.stats`
- `StatsModels`
- `scikit-learn`

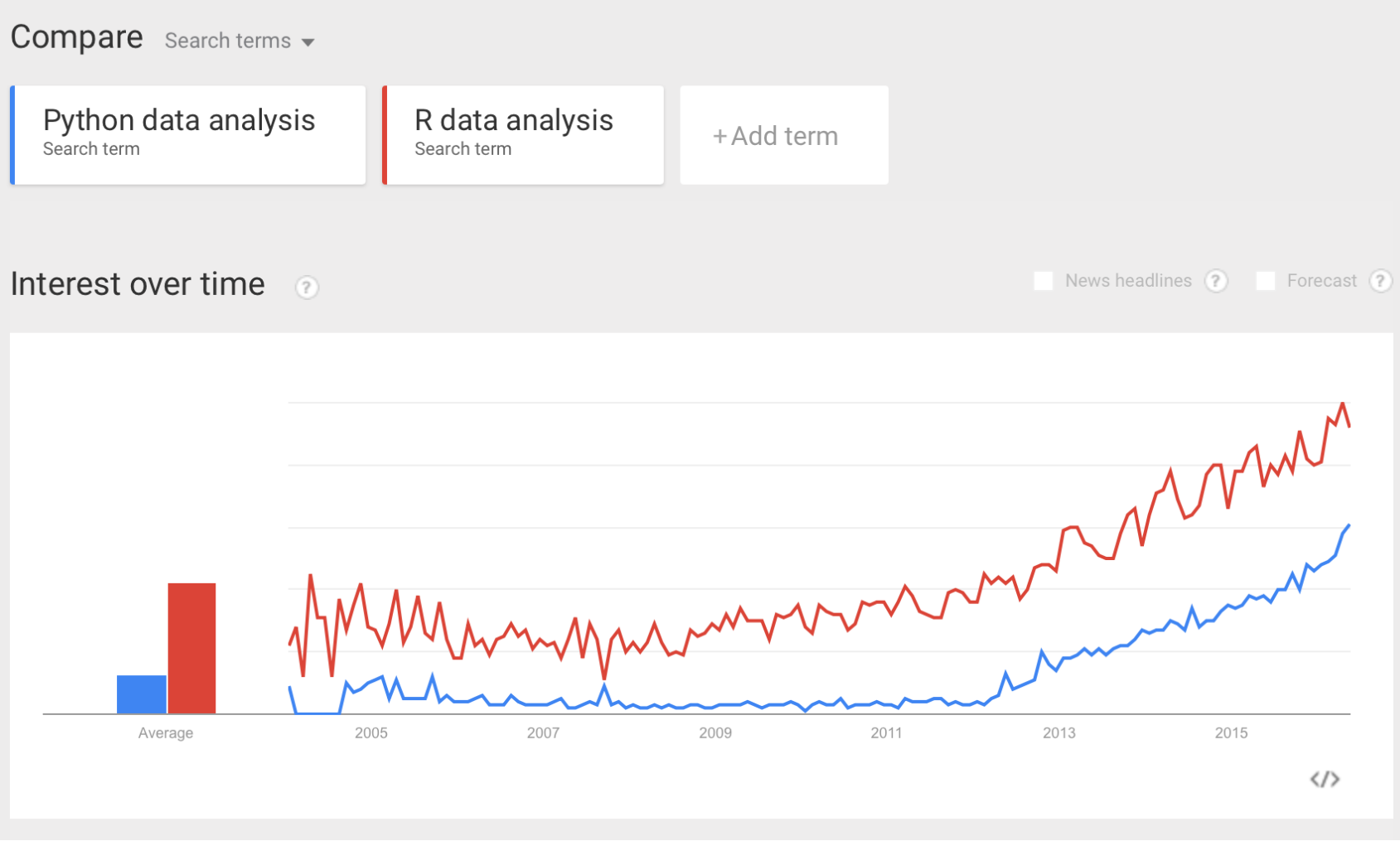
Python is gaining popularity for doing data analysis

PYPL Popularity of Programming Language Index, Worldwide



Python is gaining popularity for doing data analysis

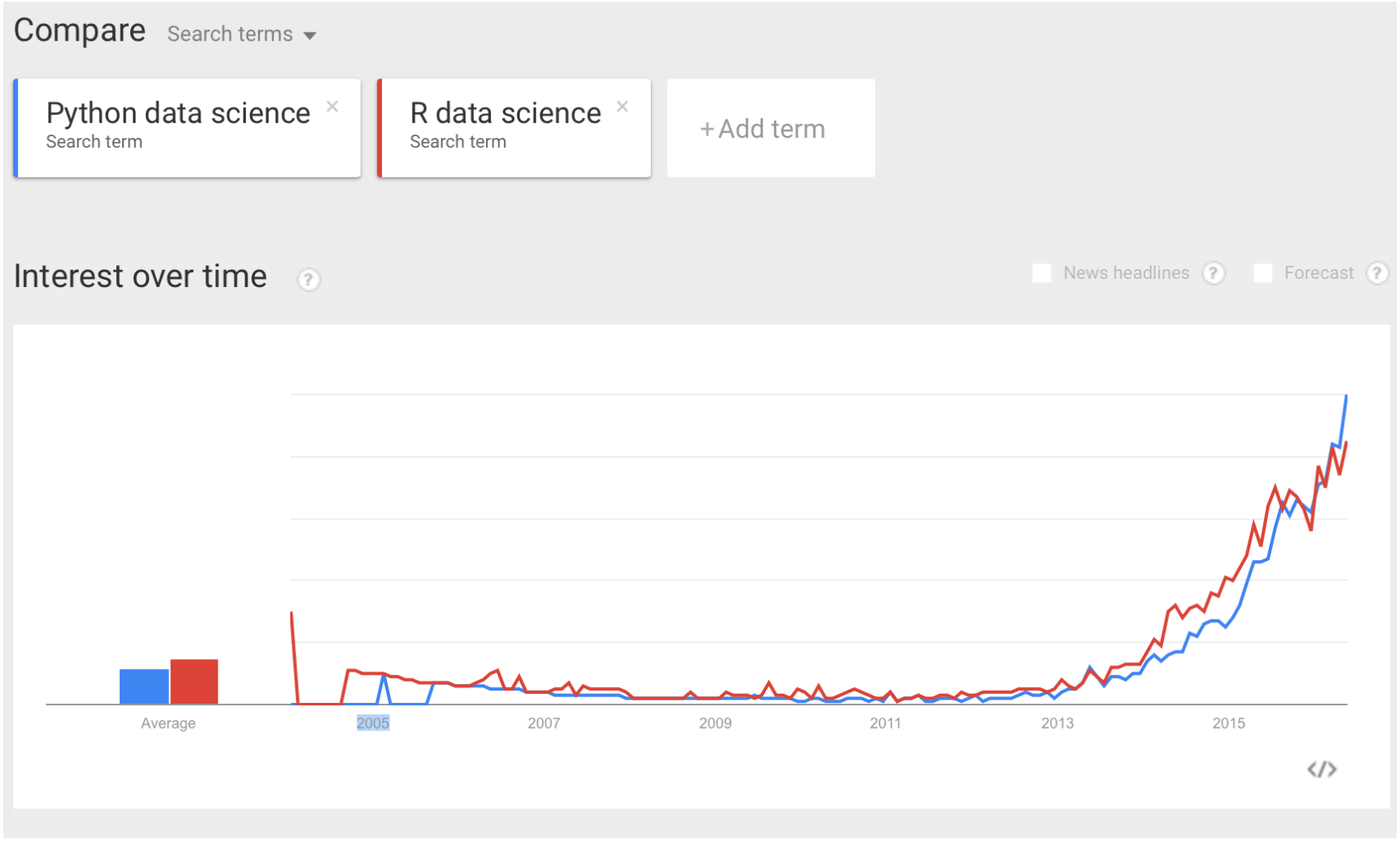
Google trends on May 22, 2016



Keyword: data analysis

Python is gaining popularity for doing data analysis

Google trends on May 22, 2016



Keyword: data science

Outline

Student evaluations of teachers (SET) are used to

- Quantify teaching effectiveness
- Compare instructors across courses
- Make hiring, firing, and promotion decisions

Are SET a valid measure of teaching effectiveness?

No!

We reanalyzed data from MacNell et al. [2014].

- Students randomized to 4 online sections of a course
- In two sections, the TAs swapped identities
- Female-identified TA was rated lower on average in all categories

Neyman-Rubin model, generalized

Student i is represented by a ticket with 4 numbers, their response to each "treatment."

$$r_{ijk} = \text{ SET given by student } i \text{ to instructor } j \text{ when they appear to have gender } k$$

$$i = 1, \ldots, N; \qquad j = 1, 2; \qquad k \in \{\text{male, female}\}$$

Numbers are fixed; randomization reveals one of the numbers.

Assume non-interference: each student's response depends only on that student's treatment.

If gender doesn't matter,

$$r_{ij\text{male}} = r_{ij\text{female}}.$$

```
# initialize PRNG
rs = np.random.RandomState(seed=1)
reps=10**5

ratings = macnell2014()

# Ratings vs reported instructor gender (difference in means)
(p, t, distr) = stratified_two_sample(ratings['overall'][ratings.taidgender==1],
ratings['overall'][ratings.taidgender==0],
                        ratings['tagender'][ratings.taidgender == 1],
                        ratings['tagender'][ratings.taidgender == 0],
                        alternative = "two-sided", stat='mean', seed = rs,
                        reps = reps, keep_dist = True)
print 'Overall rating:'
print 'Difference in means:', t
print 'P-value (two-sided):', np.round(p, 5), "\n"
```

Results

Characteristic	M-F	perm P	t-test P
Overall	0.47	0.12	0.128
Caring	0.52	0.10	0.071
Consistent	0.47	0.21	0.045
Enthusiastic	0.57	0.06	0.112
Fair	0.76	0.01	0.188
Feedback	0.47	0.16	0.054
Helpful	0.46	0.17	0.049
Knowledgeable	0.35	0.29	0.038
Praise	0.67	0.01	0.153
Professional	0.61	0.07	0.124
Prompt	0.80	0.01	0.191
Respectful	0.61	0.06	0.124
Responsive	0.22	0.48	0.013

salt and mortality

NSGK IRR stuff Millman et al. [2016]

Outline

Reproducibility crisis:

- Why Most Published Research Findings Are False (Ioannidis, 2005)
- 30–50%TO DO: of studies fail to replicate (TO DO: CITE)

Why?

- File drawer problem
- Publication bias: positive findings are more likely to get published
- P-hacking and trying many models before reporting one
- Inappropriate statistical tests

Randomization inference may ameliorate the last problem

Download `permute`!

Permutation tests and confidence sets

build passing coverage 99%

Permutation tests and confidence sets for a variety of nonparametric testing and estimation problems, for a variety of randomization designs.

- **Website (including documentation):** http://statlab.github.io/permute
- **Mailing list:** http://groups.google.com/group/permute
- **Source:** https://github.com/statlab/permute
- **Bug reports:** https://github.com/statlab/permute/issues

Installation from binaries

```
$ pip install permute
```

https://github.com/statlab/permute

Collaborators

Jarrod Millman
jarrodmillman

Philip B. Stark
pbstark

Stefan van der Walt
stefanv

References

L. MacNell, A. Driscoll, and A. N. Hunt. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, pages 1–13, 2014.

K. J. Millman, P. B. Stark, K. Ottoboni, and Naomi A. P. Stark. A case study in reproducible applied statistics: Is tagging of therapist-patient interactions reliable? Technical report, University of California, Berkeley, 2016. URL `https://github.com/statlab/nsgk`.