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

1 Introduction

2 Examples
- Gender bias in teaching evaluations
- Inter-rater reliability

3 The role of software development in Statistics

History of randomization inference - fisher - neyman model

R has several packages for randomization inference.

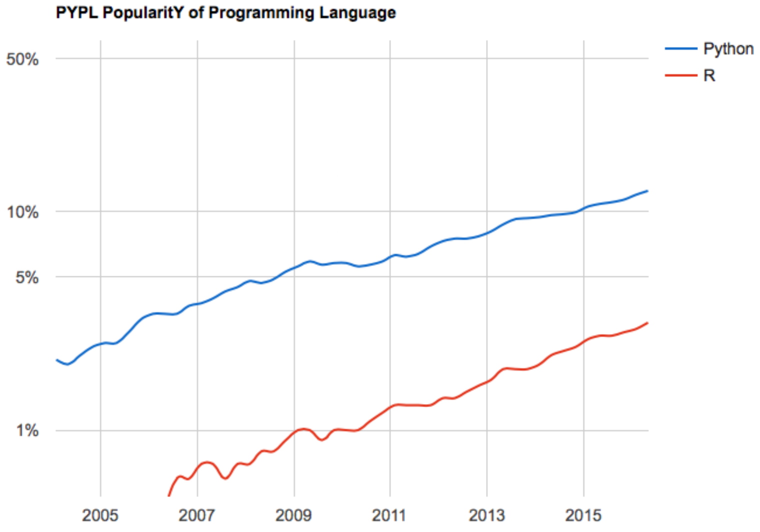
- `ri`
- `RItools`
- `coin`
- `perm`

In Python, statistics packages are limited.

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

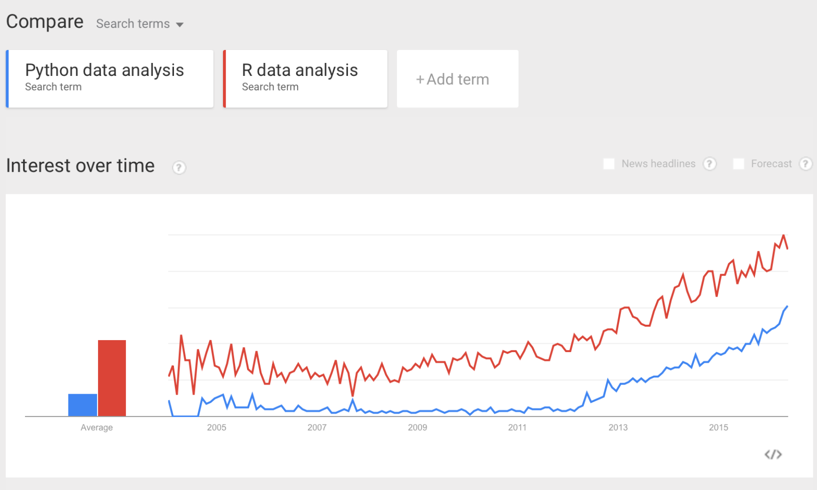
Python is gaining popularity for doing data analysis

PYPL Popularity of Programming Language Index, Worldwide



Python is gaining popularity for doing data analysis

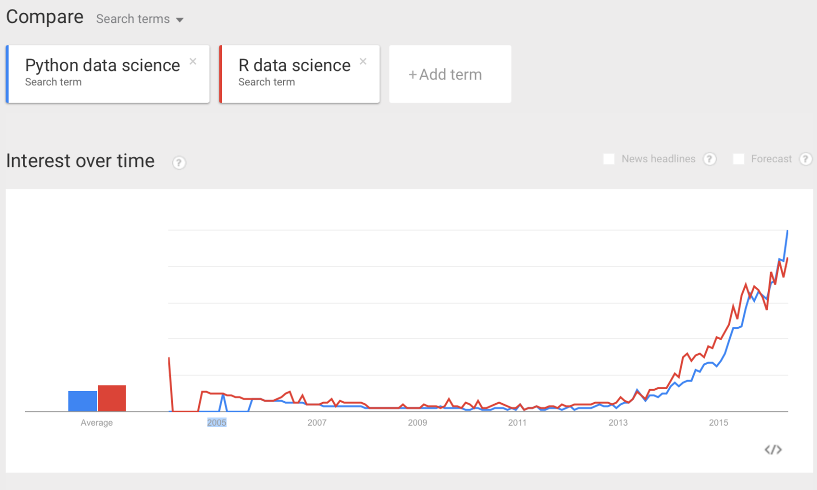
Google trends on May 22, 2016



Keyword: data analysis

Python is gaining popularity for doing data analysis

Google trends on May 22, 2016



Keyword: data science

Python for teaching Statistics

Data Science 8, Spring 2016 at UC Berkeley

Outline

Teaching Evaluations

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 were randomized to 4 online sections of a course.
- In two sections, the TAs swapped identities.
- Was the TA who identified as female rated lower on average?

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,

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

Randomization

Under this null, it's as though

1. N_m students are randomly assigned to the male TA, N_f to the female TA
2. N_{mm} of the N_m students and N_{fm} of the N_f students are told that their TA is male

We know what the data would have been for all

$$\binom{N_m}{N_{mm}} \times \binom{N_f}{N_{fm}}$$

equally likely assignments of students to sections.

This determines the conditional null distribution of **any statistic**.

Stratified two-sample test

```
# load packages
import numpy as np
from permute.data import macnell2014
from permute.stratified import stratified_two_sample

# initialize PRNG
rs = np.random.RandomState(seed=1)

# load the data
ratings = macnell2014()

# Ratings vs reported instructor gender (difference in means)
(p, t) = stratified_two_sample(group=ratings.taidgender,
                               response=ratings.overall,
                               condition=ratings.tagender,
                               alternative='two-sided',
                               stat='mean', reps=10**5)
```

Results

TO DO: UPDATE P-VALUES In all categories, the male-identified TA was rated higher.

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

Omnibus Test

Nonparametric combination of tests (NPC): combine individual p-values into a single omnibus test when there are many responses

Test whether **all null** hypotheses are true or **at least one alternative** is true.

Fisher's combining function:

$$X^2 = -2 \sum_{j=1}^{J} \ln(P_j)$$

$X^2 \sim \chi^2_{2J}$ if $\{P_j\}_{j=1}^{J}$ are independent and all nulls are true.

Omnibus Test

Ratings for different categories are **dependent**.

$\Longrightarrow$ Calibrate the distribution of X^2 using the permutation distributions of each individual statistic.

- Calculate the vector of observed values of test statistics (use the **same permutation** to compute each statistic)
- Apply the combining function to get a single test statistic X^2 (or other) for the permutation.
- Repeat a large number B times to find the permutation distribution of X^2.

Omnibus Test

```
# Initialize placeholders
ind = 0
test_distr = np.zeros((10**5, len(categories)))
pvalues = np.zeros(len(categories) )

# Loop over rating categories
for col in categories:
    (p, t, distr) = stratified_two_sample(
                                group=ratings.taidgender,
                                response=ratings[col],
                                condition=ratings.tagender,
                                alternative='two-sided',
                                stat='mean', seed = seed,
                                reps = 10**5, keep_dist = True)
    ind += 1
    test_distr[:,ind] = distr; pvalues[ind] = p

# NPC
omnibus_pvalue = npc(pvalues, test_distr, combine="fisher",
                           alternatives="two-sided")
```

Result: $P = 0$

NSGK Data

- Naomi Stark and Gilbert Kliman (NSGK) collected videos of therapy sessions with children on the autism spectrum
- A team of trained raters watched and tagged each 30-second interval of video from a collection of 183 clinically relevant tags
- Is tagging of therapist-patient interactions reliable? (Millman et al. [2016])

NSGK Data

NSGK	**IRR**
183 types of activity	T tags
8 videos	S strata
40 segments/videos	N_s items/strata
10 raters	R raters

Table : Mapping between terms from our motivating problem (NSGK) and the terms used in our general algorithm (IRR).

We observe $\{L_{s,i,c,r}\}$ for $s = 1, \ldots, S$; $i = 1, \ldots, N_s$; $c = 1, \ldots, C$; and $r = 1, \ldots, R$.
The test statistic within stratum s is

$$\rho_s \equiv \frac{1}{N_s \binom{R}{2}} \sum_{i=1}^{N_s} \sum_{r=1}^{R-1} \sum_{v=r+1}^{R} \mathbf{1}(L_{s,i,r} = L_{s,i,v})$$

$$= \frac{1}{N_s R(R-1)} \sum_{i=1}^{N_s} (y_{si}(y_{si} - 1) + (R - y_{si})(R - y_{si} - 1)).$$

Outline

1 Introduction

2 Examples
- Gender bias in teaching evaluations
- Inter-rater reliability

3 The role of software development in Statistics

Reproducibility

Monkey Cage

Does social science have a replication crisis?

Cancer Research Is Broken

There's a replication crisis in biomedicine—and no one even knows how deep it runs.

By Daniel Engber

SundayReview

Why Do So Many Studies Fail to Replicate?

Gray Matter

By JAY VAN BAVEL MAY 27, 2016

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration[*,†]

NATURE | EDITORIAL

Reality check on reproducibility

POLICY & ETHICS

Is There a Reproducibility Crisis in Science?

PLOS | MEDICINE

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • http://dx.doi.org/10.1371/journal.pmed.0020124

About 40% of economics experiments fail replication survey

By John Bohannon | Mar. 3, 2016 , 2:00 PM

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

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

Isn't data science just statistics?

If we want to

- facilitate reproducible research,
- enable people to easily use the methods we develop, and
- influence the way people do statistics more broadly,

then **we** have to build the tools.

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`.