From Paper to Program: Challenges of Implementing Permutation Tests

Kellie Ottoboni

Department of Statistics, UC Berkeley
Berkeley Institute for Data Science

June 11, 2018
ISNPS, Salerno

University of California, Berkeley
Department of Statistics

BERKELEY INSTITUTE FOR DATA SCIENCE

Outline

Introductory Statistics

Important concepts: sampling distribution, p-value, confidence intervals

Get obscured by

- Z tests, t tests
- Assumptions
- Endless formulas

Introductory Statistics

What if we could teach the concepts without the particular details?

Tools:

1. Resampling methods
2. Computers

Introductory Statistics

Permutation tests clarify concepts

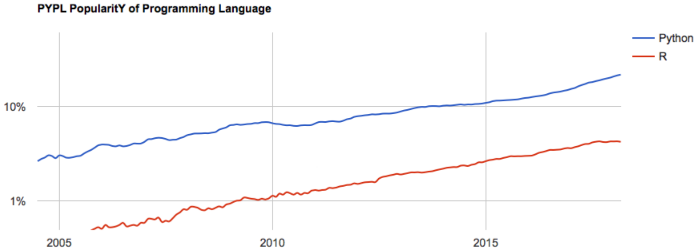
- General: it's a procedure, not a formula
- Discrete: counting instead of integration
- Design-based: assumptions come from the data collection

Hesterberg [2015]

Python

Python is gaining popularity for doing data analysis.

- General purpose language with "batteries included"
- Popular for a variety of scientific applications



Python for teaching Statistics

Data Science 8, Spring 2016 at UC Berkeley

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?

Teaching evaluations

In Boring et al. [2016], we reanalyzed data from MacNell et al. [2014].

- Students were randomized to 4 online sections of a course.
- In two sections, the instructors swapped identities.
- Was the instructor 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

Conceptually, there are two levels of randomization:

1. N_m students are randomly assigned to the male instructor, and the remaining N_f get the female instructor.
2. Of the N_j assigned to instructor j, N_{jm} are told that the instructor is male, for $j = 1, 2$.

All $\binom{N_m}{N_{mm}} \times \binom{N_f}{N_{fm}}$ assignments of students to sections are equally likely.

Stratified two-sample test

- For each instructor, permute perceived gender assignments
- Use difference in mean ratings for female-identified minus male-identified

Code

```python
# 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.tagender,
                               response=ratings.overall,
                               condition=ratings.taidgender,
                               alternative="two-sided",
                               stat="mean", reps=10**5)
```

Results

In all categories, the male-identified instructor 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 to accept **all null** hypotheses or reject **at least one alternative**

Fisher's combining function

Let $\{P_j\}_{j=1}^J$ be p-values for J hypotheses. Define

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

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

Omnibus Test

Ratings by the same student for different categories are **dependent**.

$\Longrightarrow$ Treat all ratings from a student as a vector and calibrate the distribution of X^2 using the this permutation distribution.

NPC Permutation Procedure

1. Calculate the vector of test statistics (use the **same permutation** of section memberships to compute all statistics), repeat a large number B times
2. Compute the p-value for each individual variable in each permutation relative to the other values in the distribution
3. Apply the combining function to each vector of p-values.

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.tagender,
                        response=ratings[col],
                        condition=ratings.taidgender,
                        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")
```

Conclusions

Omnibus test: $P = 0$

Reject the null hypothesis that there is no difference in ratings for any category.

$\implies$ SET measure something other than teaching effectiveness.

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

References

Anne Boring, Kellie Ottoboni, and Philip B. Stark. Teaching evaluations (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 2016. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

James V. Bradley. *Distribution-free statistical tests*. Prentice-Hall, 1968.

Ronald A. Fisher. *Design of Experiments*. New York: Hafner, 1935.

Tim C Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, 2015.

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`.

Bin Yu. Let us own data science. Institute of Mathematical Statistics (IMS) Presidental Address, ASC-IMS Joint Conference, Sydney, July 2014. URL `https://www.stat.berkeley.edu/~binyu/ps/papers2014/IMS-pres-address14-yu.pdf`.