permuter: An R Package for Randomization Inference

Kellie Ottoboni

Department of Statistics, UC Berkeley
Berkeley Institute for Data Science

June 28, 2016
useR! Stanford

University of California, Berkeley
Department of Statistics

BIDS
Berkeley Institute for Data Science

Outline

Permutation tests

- Fisher [1935] introduced permutation tests for randomized experiments
- Rely on assumptions about randomization or exchangeability, rather than parametric assumptions, IID sampling, etc.

James Bradley [1968]

"[a] corresponding parametric test is valid only to the extent that it results in the same statistical decision [as the randomization test]."

Permutation tests

R has several packages for randomization inference.

- `ri`
  "This package provides a set of tools for conducting exact or approximate inference for randomized experiments of arbitrary design. The primary functionality of the package is in the generation, manipulation and use of permutation matrices implied by given experimental designs..."
- `RItools`
  "The RItools package implements useful functions for implementing randomization inference based statistical tests. The package provides tools for testing balance of observed covariates in observational studies using the methodology of:...The package also provides outcome analysis of simple or block randomized trials (or matched observational studies) based on user defined models and test statistics."
- `coin`
  The R package coin implements a unified approach to permutation tests providing a huge class of independence tests for nominal, ordered, numeric, and censored data as well as multivariate data at mixed scales. Based on a rich and flexible conceptual framework that embeds different permutation test procedures into a common theory, a computational framework is established in coin that likewise embeds the corresponding R functionality in a common S4 class structure with associated generic functions.
- `perm`
  The package has three main functions, to perform linear permutation tests. These tests are tests where the test statistic is the sum of the product of a covariate (usually group indicator) and the scores.

TO DO: FLESH OUT INTRO!!

Outline

1 Introduction

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

3 The role of software development in Statistics

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

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.

This determines the conditional null distribution of **any statistic**.
e use the difference in mean ratings.

Stratified two-sample test

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

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**.

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

TO DO: CHECK THAT THIS IS LEGIBLE

NPC Permutation Procedure

1. Calculate the vector of observed values of test statistics (use the **same permutation** of section memberships to compute all statistics)
2. Apply the combining function to get a single combined statistic for the permutation.
3. Repeat a large number B times to find the permutation distribution of the combined statistic.

Omnibus Test

Conclusions

Omnibus test: $P = 0$

- Reject the null hypothesis that there is no difference in ratings for any category
- The male-identified instructor was rated significantly higher than the female-identified instructor on several dimensions, **even on objective measures** such as how promptly assignments were returned
- SET measure something other than teaching effectiveness

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])? Which tags do raters agree on?

Inter-rater reliability test

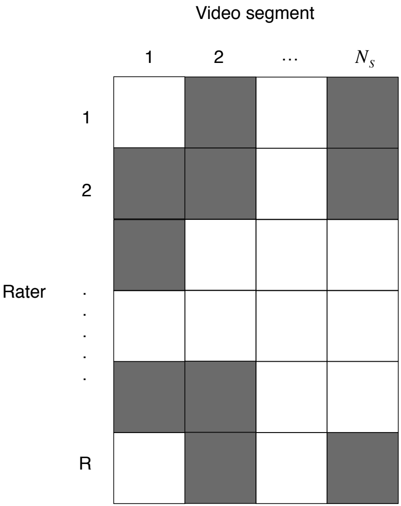
There are four dimensions. Can we simplify?

- Consider each clinical tag individually
- Do a partial hypothesis test for each video, then combine using NPC

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

Inter-rater reliability test

Is agreement within columns better than expected by chance?



Inter-rater reliability test

Define

- $\{L_{s,i,r}\}$ = indicator for whether rater r tagged item i in stratum s
- $y_{si} = \sum_{r=1}^{R} L_{s,i,r}$ = number of raters who tagged item i in stratum s

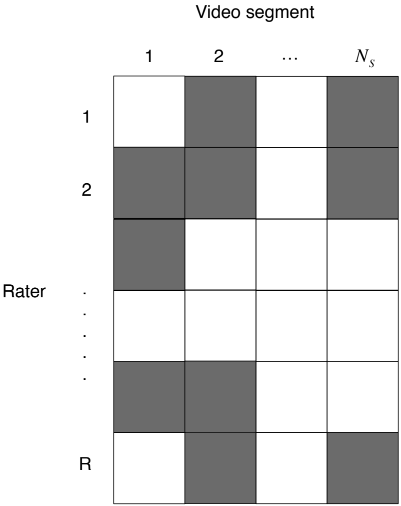
The test statistic within stratum s is

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

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

Inter-rater reliability test

Now we have a measure of concordance. What is the chance model?



Permutation test

If tags are assigned completely at random, then

- any of 2^{N_s} assignments of tags are equally likely for each rater.
- raters assign tags independently of each other

Each rater may have different "propensity" to assign a tag

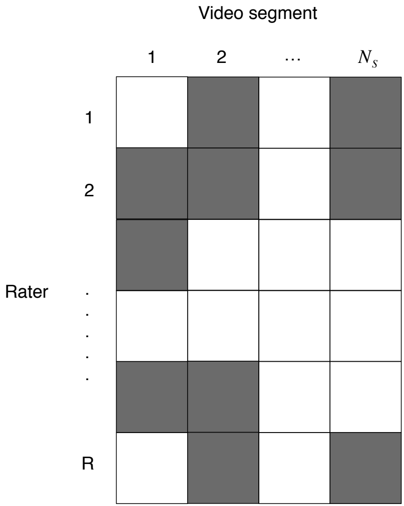
- **Solution:** condition on the number of items that a rater tagged.
- **Implied randomization:** Permute tags within rows, independently across rows and across strata.

For overall test of tag, combine using NPC:

$$T = -\sum_{s=1}^{S} \frac{P_s}{\sqrt{N_s}}$$

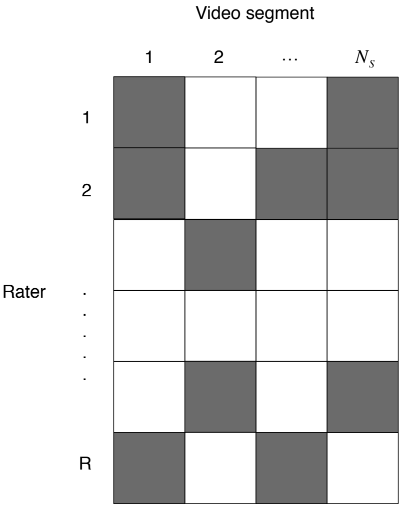
Inter-rater reliability test

Observed tags for stratum s



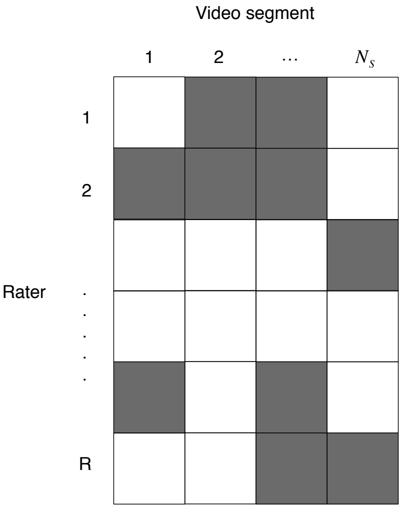
Inter-rater reliability test

Equally likely tags for stratum s, under the null



Inter-rater reliability test

Equally likely tags for stratum s, under the null



Code

Results

- 60 tags had $P < 0.05$
- Statistical vs practical significance – consult domain scientists
- Is there a more useful summary statistic than ρ_s?

Outline

Reproducibility

Why should Statisticians worry about writing software?

- Ethics
- Impact

Ethics

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

Ethics

Much of the reproducibility crisis can be traced back to bad statistics.

- Publication bias: positive findings are more likely to get published
- P-hacking and the garden of forking paths (Gelman and Loken [2013])
- Inappropriate statistical tests (Randomization inference may help here)

It is our responsibility to make it easy for researchers to do the right statistics.

Impact

Let us own data science (Yu [2014]).

If we want to

- facilitate reproducible scientific research,
- enable people to use the methods we develop (correctly!), and
- influence the way people do statistics more broadly,

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

Download `permuter`!

https://github.com/statlab/permuter

Collaborators

Jarrod Millman
jarrodmillman

Philip B. Stark
pbstark

References

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

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

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished paper, 2013.

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`.