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

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

Permutation tests

- Fisher [1935] introduced permutation tests for the analysis of randomized experiments
- They rely on assumptions about randomization or exchangeabilityto generate the distribution of a test statistic under the null hypothesis
- They were largely replaced by parametric hypothesis tests because they were computationally infeasible, but are making a comeback

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."

Teaching Statistics with Permutation Tests

Data Science 8, Spring 2016 at UC Berkeley

Teaching Statistics with Permutation Tests

Goals for `permuter`:

- Create a comprehensive package for permutation tests
  - Building blocks of permutation tests
  - Transparent, not a black box, to facilitate understanding
- Write code with algorithms and computation in mind
- Compile examples and case studies alongside the package

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?

Teaching evaluations

In **?**, 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

```r
# Load data
library(permuter)
data(macnell)


# Calculate the observed test statistic (difference in means)
tst_statistic <- function(x) {
  mean(x[macnell$taidgender == 0], na.rm = TRUE) -
  mean(x[macnell$taidgender == 1], na.rm = TRUE)
  }
observed <- tst_statistic(macnell$overall)

# Test overall ratings vs reported instructor gender
distr <- stratified_two_sample(response = macnell$overall,
                               group = macnell$taidgender,
                               stratum = macnell$tagender,
                               stat = "mean", reps = 1e5)


# Calculate two-sided p-value
p <- t2p(observed, distr, alternative = "two-sided")
```

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
categories <- colnames(macnell)[2:16]
test_distr <- matrix(0, nrow = 1e5, ncol = length(categories))
pvals <- rep(0, length(categories))
colnames(test_distr) <- names(pvals) <- categories

# Loop over rating categories
for(col in categories){
  set.seed(678)
  observed <- tst_statistic(macnell[ ,col])
  distr <- stratified_two_sample(response = macnell[ ,col],
                                 group = macnell$taidgender,
                                 stratum = macnell$tagender,
                                 stat = "mean", reps = 1e5)
  test_distr[ ,col] <- distr
  pvals[col] <- t2p(observed, distr, alternative = "two-sided")
}

# NPC
omnibus_pvalue <- npc(pvals, 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.

Inter-rater reliability test

- 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])?

Inter-rater reliability test

There are four dimensions. Can we simplify?

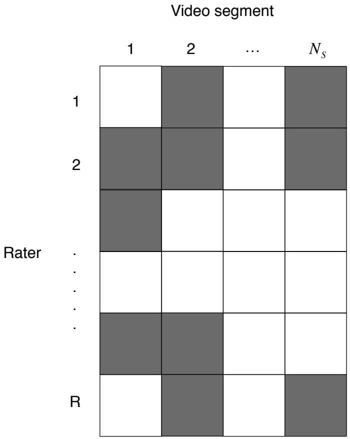
- Consider each clinical tag individually
- Do a partial hypothesis test for each video, then combine using NPC

NSGK	**IRR**
183 types of activity	T tags
8 videos	S strata
$\sim$35 segments/video	N_s items/stratum
10 raters	R raters

Inter-rater reliability test

We need

- a test statistic for concordance
- a chance model



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

$$\begin{aligned}\rho_s &\equiv \frac{1}{N_s\binom{R}{2}} \sum_{i=1}^{N_s} \sum_{r=1}^{R-1} \sum_{v=r+1}^{R} \mathbf{1}(L_{s,i,r} = L_{s,i,v}) \\ &= \frac{1}{N_s R(R-1)} \sum_{i=1}^{N_s} (y_{si}(y_{si}-1) + (R-y_{si})(R-y_{si}-1)).\end{aligned}$$

Permutation test

If tags are assigned completely at random, conditional on the number of items each rater tagged, then

- all possible permutations of tags are likely for an individual rater
- raters assign tags independently of each other

IRR Permutation Test

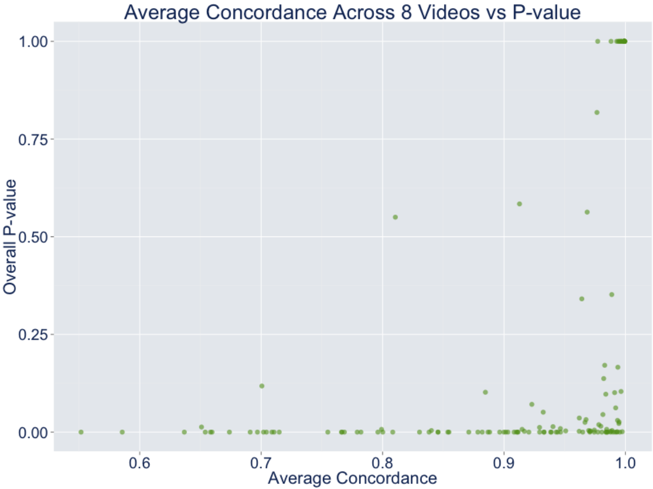
- Compute ρ_s for $s = 1, \ldots, S$ for the observed data
- Permute tags within rows, independently across rows and across strata, to get the permutation distributions
- For the observed and permutation ρ_s values, find the corresponding p-value P_s for each permutation
- The NPC test statistic is $T = -\sum_{s=1}^{S} \frac{P_s}{\sqrt{N_s}}$.

Code

```r
library(permuter)
data(nsgk)
time_stamps <- c(36, 32, 35, 37, 31, 35, 40, 32)

# nsgk_mat is a list of lists:
# First level indexes the tag, second level indexes the video
# See the vignette nsgk.Rmd for details data(nsgk) --> nsgk_mat

# Example of how we analyze a single tag.
tag1 <- nsgk_mat[[1]]

# Find the distribution of the IRR test statistic for each video
results_by_video <- lapply(tag1, function(x){
    res <- irr_ts_distribution(x, reps = 1000, keep_dist = TRUE,
                               seed = 101)
    })
tag_distr     <- sapply(results_by_video, function(x) x$dist)
tag_pvals  <- sapply(results_by_video, function(x) x$pvalue)

# Combine across videos with NPC
tag_npc_res <- irr_npc_distribution(tag_distribution,
                         size = time_stamps, pvalues = tag_pvals)
```

Results

- 93 tags had $P < 0.05$
- Statistical vs practical significance – consult domain scientists



Outline

Software and Statistics

Why should Statisticians worry about writing software?

- Ethics
- Impact

Ethics

Cancer Research Is Broken

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

By Daniel Engber

Monkey Cage

Does social science have a replication crisis?

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

Impact

Bin Yu [2014]

Let us own data science.

Statisticians have to build the tools to

- facilitate reproducible scientific research,
- enable people to use the methods we develop (correctly!), and
- impact the way people think about data.

Collaborators

Jarrod Millman
jarrodmillman

Philip B. Stark
pbstark

`https://github.com/statlab/permuter`

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`.