Lecture 12: Permutation tests II

Pratheepa Jeganathan

04/29/2019

Recall

- One sample sign test, Wilcoxon signed rank test, large-sample approximation, median, Hodges-Lehman estimator, distribution-free confidence interval.
- Jackknife for bias and standard error of an estimator.
- Bootstrap samples, bootstrap replicates.
- Bootstrap standard error of an estimator.
- Bootstrap percentile confidence interval.
- Hypothesis testing with the bootstrap (one-sample problem.)
- Assessing the error in bootstrap estimates.
- Example: inference on ratio of heart attack rates in the aspirin-intake group to the placebo group.
- The exhaustive bootstrap distribution

- Discrete data problems (one-sample, two-sample proportion tests, test of homogeneity, test of independence)
- Two-sample problems (location problem - equal variance, unequal variance, exact test or Monte Carlo, large-sample approximation, H-L estimator, dispersion problem, general distribution)
- Permutation tests (permutation test for continuous data, different test statistic, accuracy of permutation tests)

Permutation test (discrete data)

Example (The lady tasting tea)

- This example is from `The Design of Experiments` by Fisher (1935), chapter II link here.
- A British lady claimed that she can tell whether the tea or the milk was added first to a cup.
- Fisher proposed a randomized experiment.
- The null hypothesis is that the lady has no ability to taste the difference.

Example (Lady tasting tea)

- Experiment:
  - The lady provided with 8 randomly ordered cups of tea.
    - In four cups, tea was added first.
    - In other four cups, milk was added first.
  - The lady has to select 4 cups prepared by one method.
  - The lady knew the method used for the experiment.
- Test statistic $T =$ the number of successes in selecting the 4 cups (the number of cups of the given type successfully selected)
- What is the distribution of the test statistic T under H_0?

Example (Lady tasting tea)

- The distribution of T under H_0 can be computed using the number of permutations because the judgement are equally likely.
- Using the combination formula, $n = 8$ and $k = 4$, there are $\binom{8}{4} = 70$ possible combinations.

Example (Lady tasting tea)

# of success	Arrangement	# of permutations
0	0000	$\binom{4}{0} \times \binom{4}{4} = 1$
1	x000,0x00,00x0,000x	$\binom{4}{1} \times \binom{4}{3} = 16$
2	xx00,x0x0,x00x,0xx0,0x0x,00xx	$\binom{4}{2} \times \binom{4}{2} = 36$
3	xxx0,xx0x,x0xx,0xxx	$\binom{4}{3} \times \binom{4}{1} = 16$
4	xxxx	$\binom{4}{4} \times \binom{4}{0} = 1$

- The number of success T is distributed according to the hyper geometry distribution under the null hypothesis.
- $P(T = t) = \frac{\binom{4}{t}\binom{4}{4-t}}{\binom{8}{4}}$.

Example (Lady tasting tea)

```r
library(dplyr)
t = 0:4
hypergeometry = (choose(4,t)*choose(4,4-t))/choose(8,4)
df = data.frame(t=t, p.t = round(hypergeometry, digits=3));
```

```
##   t   p.t
## 1 0 0.014
## 2 1 0.229
## 3 2 0.514
## 4 3 0.229
## 5 4 0.014
```

- The critical region for rejection of the null the lady has no ability to taste the difference at 5% significance level is the single case of 4 successes of 4 possible. That is, $T \geq 4$.

Example (Lady tasting tea)

- If the lady distinguish all the cups correctly only was Fisher willing to reject the null hypothesis (with 8 cups) at .014 significance level.

Example (Lady tasting tea)

- If n is large, we can use Monte Carlo method to approximate the p-value.

```
observed = c("milk", "milk", "milk", "milk", "tea",
  "tea", "tea", "tea")
t.0 = sum(observed[1:4]=="milk");t.0
```

```
## [1] 4
```

```r
nperm = 10000
permutations = replicate(nperm, sample(8, replace = FALSE))
matches = apply(permutations, 2, function(i){
  sum(observed[i][1:4] == "milk")
  })
data.frame(t=t,
           p.t = round(hypergeometry, digits=3),
           monte = round(table(matches)/nperm, digits=3))
```

```
##   t   p.t monte.matches monte.Freq
## 1 0 0.014             0      0.014
## 2 1 0.229             1      0.233
## 3 2 0.514             2      0.512
## 4 3 0.229             3      0.226
## 5 4 0.014             4      0.014
```

- P-value is $P(T \geq t_0) = P(T \geq 4) = .014$ so reject the null hypothesis.

- If we increase the number of cups to 16

```
observed = c("tea", "milk", "milk", "milk", "tea",
  "tea", "tea", "milk", "tea",
  "milk","tea","tea", "tea",
  "milk","milk","milk")
nperm = 10000
permutations = replicate(nperm, sample(16,
  replace = FALSE))

matches = apply(permutations, 2,
  function(i){
  sum(observed[i][1:8] == "milk")
  })
```

```r
data.frame(monte = round(table(matches)/nperm,
  digits=3))
```

```
##   monte.matches monte.Freq
## 1             0      0.000
## 2             1      0.005
## 3             2      0.061
## 4             3      0.235
## 5             4      0.384
## 6             5      0.244
## 7             6      0.065
## 8             7      0.006
## 9             8      0.000
```

- If the lady tasted 16 cups, it would be possible to reject H_0 without requiring perfect judgement.

Fisher's exact test (recall from lecture on discrete data problem)

- Discrete data problem.

		Truth		
		Milk	Tea	
Guess	Milk	4	0	**4**
	Tea	0	4	**4**

- Testing of two probabilities/testing association of two discrete variables when the marginals are fixed.
- The exact p-value is computed using the hyper geometry distribution (Fisher).

```r
df = data.frame(milk=c(4,0), tea = c(0,4))
fisher.test(df, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df
## p-value = 0.01429
## alternative hypothesis: true odds ratio is greater than
## 95 percent confidence interval:
##  2.003768      Inf
## sample estimates:
## odds ratio
##        Inf
```

- Reject H_0.

```r
df = data.frame(milk=c(3,1), tea = c(1,3))
fisher.test(df, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df
## p-value = 0.2429
## alternative hypothesis: true odds ratio is greater than
## 95 percent confidence interval:
##  0.3135693       Inf
## sample estimates:
## odds ratio
##   6.408309
```

- Do not reject H_0.

Exchangeability

Exchangeability

- A sufficient condition for permutation test is exchangeable of observations.
  - Consider random sample $X_1, \cdots X_n$.
  - If their joint distribution are equal under permutations Π

    $$P_{X_1,\cdots X_n}(x_1, \cdots, x_n) = P_{X_{\Pi(1)},\cdots X_{\Pi(n)}}\left(x_{\Pi(1)}, \cdots, x_{\Pi(n)}\right),$$

    then $X_1, \cdots X_n$ is exchangeable.
- This is a weaker assumption than independence of observations.
- An infinite sequence $X_1, \cdots, X_n, \cdots$ is said to be exchangeable if for all $n = 2, 3, \cdots$,

  $$X_1, \cdots X_n \overset{d}{=} X_{\Pi(1)}, \cdots, X_{\Pi(n)}$$

  for all $\Pi \in S(n)$, where $S(n)$ is the group of permutations of $\{1, 2, \cdots, n\}$.

Example (Exchangeability)

- Independent and identically distributed observations are exchangeable.
  - If $X_1, \cdots, X_n$ are independent and identically distributed, they are exchangeable, but not conversely.
- Samples without replacement from a finite population are exchangeable:
  - An urn contains b black balls, r red balls, y yellow balls, and so forth.
  - A series of balls are extracted from the urn.
  - After the i-th extraction, the color of the ball X_i is noted and k balls of the same color are added to the urn, where k can be any integer, positive, negative, or zero.
  - The set of random events $\{X_i\}$ form an exchangeable sequence, but not independent.

Example (Exchangeability)

- $\boldsymbol{X} = [X_1, \cdots, X_n]^T \sim \mathrm{MVN}(\boldsymbol{\mu}, \Sigma)$, $\Sigma = \begin{pmatrix} \sigma & \rho & \cdots & \rho \\ \vdots & & & \\ \rho & \rho & \cdots & \sigma \end{pmatrix}$, $\boldsymbol{X}$ is exchangeable, MVN stands for multivariate normal distribution.
- A simple transformation will ensure that observations are exchangeable.
  - Suppose X comes from a population with mean μ and distribution $F(t-\mu)$.
  - Y comes from a population with mean ν and distribution $F(t-\nu)$ and independent of X.
  - Define $X^{'} = X - \mu$ and $Y^{'} = Y - \nu$.
  - $X^{'}$ and $Y^{'}$ are exchangeable.

Example (Exchangeability)

- Flip a coin 20 times and we know there is 17 heads and 3 tails. If the outcome of 20 flips is exchangeable, then, we don't think of the positions that the 3 tails can occupy as being special.

Exchangeability and de Finetti's Theorem

- de Finetti's theorem involves exchangeable 0-1 binary random variables $X_1, \cdots, X_n, \cdots$.
- de Finetti shows that a binary sequence $X_1, \cdots, X_n, \cdots$ is exchangeable if and only if there exists a distribution function F on $[0, 1]$ such that for all n,

$$p\left(x_1, \cdots, x_n\right) = \int_0^1 \theta^{s_n} \left(1 - \theta\right)^{n - s_n} dF\left(\theta\right),$$

  where $s_n = \sum_{i=1}^n x_i$.
  - de Finetti (1931) shows that all exchangeable binary sequences are mixtures of Bernoulli sequences.
  - Bernoulli distribution is obtained by conditioning with θ:

$$P\left(x_1, \cdots, x_n | \theta\right) = \theta^{s_n} \left(1 - \theta\right)^{n - s_n}.$$

  - $X_1, \cdots, X_n | \theta$ IID $\Rightarrow X_1, \cdots, X_n$ is exchangeable for all n.

Exchangeability and de Finetti's Theorem

- Hewitt and Savage (1955) generalized de Finetti's theorem to any infinite exchangeable sequences.
- Diaconis and Freedman (1980) generalized de Finetti's theorem to finite exchangeable sequences.

Application

Example (For modern data)

- Permutation test for autism brain imaging data (Seiler 2016): link here.
- Reference: (Nichols and Holmes 2002)(https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.1058).
  - Multiple testing: p-value adjustment using permutation method (Westfall, Young, and others 1993).

Example (neuroimaging experiments)

- Preprocessed neuroimaging data from the Autism Brain Imaging Data Exchange (ABIDE). The data is openly available on the ABIDE website (Craddock et al. 2013).
  - ABIDE is a collaboration of 16 international imaging sites.
  - Neuroimaging data from 539 individuals suffering from Autism Spectrum Disorder (ASD) and 573 typical controls.
  - In this analysis, we subset 40 participants (all acquired at Stanford).
  - Measured cortical thickness voxel-by-voxel.

Example (neuroimaging experiments)

- Test voxelwise distribution of cortical thickness in autism population and healthy controls.
  - Two-sample problem (voxelwise).
  - Use Wilcoxon rank sum test (voxelwise).
  - If we report all significant voxel at significance level of $\alpha = .05$, we will report many random results.
- Adjust p-values for multiple testing using permutation approach.

Example (neuroimaging experiments)

- Single threshold test (Nichols and Holmes 2002; Westfall, Young, and others 1993)
  - Test statistic for testing each voxel: mean difference statistic T^k, where k denotes the k-th voxel.
  - For each possible i-th resampling, compute $t_i^{\max}$, maximum of voxel statistic.
  - $t_i^{\max}$ gives the permutation distribution for $T^{\max}$.
  - Define the critical threshold is the (C+1) largest member of the permutation distribution for $T^{\max}$, where $C = [\alpha N]$, that is αN rounded down. For example, $C = [.05 \times 40] = 2$ and the threshold is $T_{(3)}^{\max}$, the third largest member.
  - Voxels with statistics exceeding this threshold $T^k \geq T_{(3)}^{\max}$ exhibit evidence against the corresponding voxel hypotheses at level $\alpha = .05$.
  - Corrected P-value for each voxel is the proportion of the permutation distribution for the maximal statistic that is greater than or equal to voxel statistic.
    - adjusted p-value$^k = \dfrac{\#\{T^{\max} \geq T^k\}}{\#\text{permutations}}$.

Summary

- A sufficient condition for permutation test is exchangeable of observations.
- If the observations are not exchangeable, then some permutations are more likely than others.
- When doing permutation tests, in order to control the probability of type I error, one must establish that the observations are exchangeable under H_0.
- Permutation approaches can be used for adjusting p-values in multiple testing problems.

References

Li:C2016: Seiler (2016). Lecture Notes on Nonparametric Statistics.

Li:H1997: Holmes (1997). Lecture Notes on Computer Intensive Methods in Statistics.

Wikipedia: Lady tasting tea.

F: 1935: The design of experiments (1935) Chapter I-III.

Craddock, Cameron, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, et al. 2013. "The Neuro Bureau Preprocessing Initiative: Open Sharing of Preprocessed Neuroimaging Data and Derivatives." *Neuroinformatics* 4.

Diaconis, Persi, and David Freedman. 1980. "Finite Exchangeable Sequences." *The Annals of Probability*. JSTOR, 745–64.

Hewitt, Edwin, and Leonard J Savage. 1955. "Symmetric Measures on Cartesian Products." *Transactions of the American Mathematical Society* 80 (2). JSTOR: 470–501.

Nichols, Thomas E, and Andrew P Holmes. 2002. "Nonparametric Permutation Tests for Functional Neuroimaging: A Primer with Examples." *Human Brain Mapping* 15 (1). Wiley Online Library: 1–25.

Westfall, Peter H, S Stanley Young, and others. 1993. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Vol. 279. John Wiley & Sons.