Permutation tests

Patrick Breheny

September 25

The conditioning idea

- In many hypothesis testing problems, information can be divided into portions that pertain to the hypothesis and portions that do not
- The usual approach is to condition on the portions which do not pertain to the null hypothesis, thereby obtaining a more powerful test that focuses on the relevant aspects of the data
- Parametric examples:
  - Comparison of Poisson rates
  - Fisher's Exact Test

Nonparametric conditioning

- In nonparametric settings, we will usually be conditioning on the observed values of the data: $\{x_i\}$
- Under the null hypothesis that all these observed values are being drawn independently from a single distribution F, all permutations of $\{x_i\}$ are equally likely
- This fact can be used to carry out *permutation tests* of null hypotheses, conditional on the observed values, without assuming anything about F

Decomposition

- Let us introduce the following notation, decomposing $\mathbf{x}$ into two parts:
    - Let $\mathbf{x}_{(\cdot)}$ denote the vector of order statistics
    - Let $\mathbf{r}$ denote the vector of ranks; *i.e.*,

$$r_i = \sum_j I(x_j \leq x_i)$$

- We will ignore the possibility of ties in this course, although be aware the above definition must be modified when ties are present
- Note that this is indeed a decomposition:
    - Given $\mathbf{x}$, we can construct $\mathbf{x}_{(\cdot)}$ and $\mathbf{r}$
    - Given $\mathbf{x}_{(\cdot)}$ and $\mathbf{r}$, we can recover $\mathbf{x}$

The three null hypotheses: Introduction

- Three null hypotheses are common in nonparametric statistics, although of course, others are possible
- We will abbreviate them H_0, H_1, and H_2

H_0

- H_0 is the hypothesis of i.i.d. data
- Specifically, $H_0 : f(\mathbf{x}) = \prod f(x_i)$
- **Theorem**: Under H_0,
  - (i) $\mathbf{r}$ and $\mathbf{x}_{(\cdot)}$ are independent
  - (ii) $\mathbb{P}(\mathbf{r} = \mathbf{r}^*) = \frac{1}{n!}$ for all $\mathbf{r}^* \in \mathcal{R}$, where $\mathcal{R}$ is the set of all permutations of $\{1, \ldots, n\}$

H_1

- H_1 is the hypothesis of symmetric i.i.d. data
- Specifically, H_1 supposes that H_0 holds and that $f(x) = f(-x)$
- To handle this null hypothesis, we alter our decomposition in the following way: let
  - Let $s_i = \text{sign}(x_i)$
  - $|\mathbf{x}|_{(\cdot)}$ the order statistics of $\{|x_1|, \ldots, |x_n|\}$
  - r_i^+ denote the rank of $|x_i|$ among $\{|x_1|, \ldots, |x_n|\}$

H_1 (cont'd)

Theorem: Under H_1,

(i) $\mathbf{s}$, $\mathbf{r}^+$, and $|\mathbf{x}|_{(\cdot)}$ are mutually independent

(ii) $\mathbb{P}(\mathbf{s} = \mathbf{s}^*) = \left(\frac{1}{2}\right)^n$, where $\mathbf{s}^*$ is any n-vector of $\{-1, 1\}$

(iii) $\mathbb{P}(\mathbf{r}^+ = \mathbf{r}^*) = \frac{1}{n!}$

H_2

- H_2 is the hypothesis of independence of X and Y
- Specifically, H_2 hypothesizes that $f(\mathbf{x}, \mathbf{y}) = \prod_i g(x_i)h(y_i)$, where g and h are arbitrary density functions
- **Theorem:** Letting $\mathbf{q}$ denote the vector of ranks of $\mathbf{y}$, the following hold under H_2:
  (i) $\mathbf{r}$, $\mathbf{q}$, $\mathbf{x}_{(\cdot)}$, and $\mathbf{y}_{(\cdot)}$ are all mutually independent
  (ii) $\mathbb{P}(\mathbf{r} = \mathbf{r}^*) = \mathbb{P}(\mathbf{q} = \mathbf{q}^*) = \frac{1}{n!}$

Test statistics

- Permutation tests, like all hypothesis tests, begin with a test statistic T
- Examples: Difference in means, t-test statistic, difference in medians
- Any statistic can be used with the permutation test, but
  - To remain sensible, T should change in a monotone fashion as the evidence against H_0 grows stronger
  - Some test statistics will have greater power than others

Achieved significance level

- Having observed $\hat{T}$, the *achieved significance level* (ASL) of the test is the probability of observing at least that large a value when H_0 is true:

$$ASL = \mathbb{P}_0\{T(\mathbf{X}^*) \geq T(\mathbf{x})\},$$

  where $\mathbf{X}^*$ follows the null distribution

- For the hypotheses defined earlier, the above null distribution is straightforward to evaluate via permutation testing because for each hypothesis, all permutations of the ranks occur with equal probability

Achieved significance level (cont'd)

- Thus, the null distribution of $T(\mathbf{X}^*)$, conditional on the order statistics, is known
- For example, under H_0, the null distribution places probability $1/n!$ on each $\mathbf{x}_{(\mathbf{r}^*)}$ for $\mathbf{r}^* \in \mathcal{R}$, so that

$$ASL = \frac{1}{n!} \sum_{\mathcal{R}} \mathbf{1}\left\{T(\mathbf{x}^*) \geq T(\mathbf{x})\right\}$$

- **Theorem:** For any sample size n, any distribution F, and any $\alpha \in (0, 1)$,

$$\mathbb{P}_0\{ASL \leq \alpha\} \leq \alpha$$

Monte Carlo integration

- The ASL is straightforward to calculate, but not easy, because $n!$ is a big number
- Thus, in practice, ASL is usually approximated via the following Monte Carlo algorithm:
  (1) Draw $\mathbf{r}^*$ randomly from $\mathcal{R}$ B times
  (2) For each $\mathbf{r}^*$, calculate $\hat{T}^* = T(\mathbf{x}_{(\cdot)}, \mathbf{r}^*)$
  (3) $\widehat{ASL} = B^{-1} \sum \mathbf{1}(\hat{T}_b^* \geq \hat{T})$
- By the law of large numbers, $\widehat{ASL}$ converges to ASL

Homework

Homework: Define the "accuracy" of a Monte Carlo approximation to be the standard error of $\widehat{ASL}$ over the true ASL.

(a) How many permutations are needed to achieve 10% accuracy when $ASL = .1$?

(b) How many permutations are needed to achieve 10% accuracy when $ASL = .01$?

Homework

- The website has data from a study which examined the driving habits of illegal drug users as compared to non-illegal drug users
- The outcome we will look at is following distance
- It may be hypothesized that drug users like to engage in risky behavior and follow at closer speeds than other drivers

Homework

Homework:

(a) Test the null hypothesis that the mean following distance of drug users is the same as that of non-illegal drug users using a t-test

(b) Test the null hypothesis that the distribution of following distance is the same in both groups using a permutation test with test statistic:

$$T = \left| \frac{\sum_i g_i x_{(r_i^*)}}{\sum_i g_i} - \frac{\sum_i (1-g_i) x_{(r_i^*)}}{\sum_i (1-g_i)} \right|,$$

where g_i is a 0-1 indicator of group membership

(c) Test the same null hypothesis using a test statistic that compares the absolute difference of medians of the two groups

(d) Briefly comment on the results of the three tests