**36-700: Probability and Mathematical Statistics** **Spring 2019**

# Lecture 14: Permutation Test, Multiple Testing

*Lecturer: Jing Lei*

## 14.1 Review and Outline

In the last class we discussed hypothesis testing:

1. The $\chi^2$ test for multinomials.
2. p-values.

Today we will discuss permutation tests and multiple testing.

## 14.2 Permutation test

Now we consider a non-parametric two-sample test, where we observe:

$$X_1, \ldots, X_m \sim F_X,$$

and

$$Y_1, \ldots, Y_n \sim F_Y.$$

The hypotheses we would like to test are:

$$\begin{aligned} H_0 : &\quad F_X = F_Y \\ H_1 : &\quad F_X \neq F_Y. \end{aligned}$$

A typical example is in a drug trial where one set of people are given a drug and the other set are given a placebo. We then would like to know if there is some difference in the outcomes of the two populations or if they are identically distributed.

There are various possible test statistics, but a common one is to use a difference in means:

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \left| \frac{1}{m} \sum_{i=1}^{m} X_i - \frac{1}{n} \sum_{i=1}^{n} Y_i \right|.$$

One could also standardize this statistic by its variance, or consider more complex test statistics based on signs and ranks. Let us denote the test statistic computed on the data we observed as $T_{\text{obs}}$.

In general, since we have not assumed anything about $F_X$ and $F_Y$ it is not easy to compute the distribution of our test statistic, and approximations (based on a CLT for instance) might be quite bad. The permutation test, gives a way to design an *exact* $\alpha$ level test without making any approximations.

The idea of the permutation test is simple. Define $N = m + n$ and consider all $N!$ permutations of the data $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$. For each permutation we could compute our test statistic $T$. Denote these as $T_1, \ldots, T_{N!}$.

The key observation is: **under the null hypothesis each value $T_1, \ldots, T_{N!}$ has the *same* distribution (even if we do not know what it is).**

Suppose we reject for large values of $T$. Then we could simply define the p-value as:

$$\text{p-value} = \frac{1}{N!} \sum_{i=1}^{N!} \mathbb{I}(T_i > T_{\text{obs}}).$$

It is important to note that this is an exact p-value, i.e. no asymptotic approximations are needed to show that rejecting the null when this p-value is less than $\alpha$ controls the Type I error at $\alpha$. Here is a toy-example from the Wasserman book:

**Example 2:** Suppose we observe $(X_1, X_2, Y_1) = (1, 9, 3)$. Let $T(X_1, X_2, Y_1)$ be the absolute difference in means, i.e. $T(X_1, X_2, Y_1) = 2$. The permutations are:

| permutation | value of $T$ |
|---|---|
| (1,9,3) | 2 |
| (9,1,3) | 2 |
| (1,3,9) | 7 |
| (3,1,9) | 7 |
| (3,9,1) | 5 |
| (9,3,1) | 5 |

We could use this to calculate the p-value by counting how often we got a larger value than 2:

$$\text{p-value} = \frac{4}{6} \approx 0.67,$$

so most likely we would not reject the null hypothesis in this case. Typically, we do not calculate the exact p-value (although in principle we could) since evaluating $N!$ test statistics would take too long even for moderately large $N$ (When $N = 100$, this number is larger than

the number of atoms in the universe.) Instead we approximate the p-value by drawing a few random permutations and using them. This leads to the following algorithm for computing the p-value using a permutation test:

**Algorithm for Permutation Test**

1. Compute the observed value of the test statistic $t_{\text{obs}} = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$.
2. Randomly permute the data. Compute the statistic again using the permuted data.
3. Repeat the previous step $B$ times and let $T_1, \ldots, T_B$ denote the resulting values.
4. The approximate p-value is

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

## 14.3 Multiple Testing

The problem of multiple testing is one that is fundamental to a lot of science. Typical modern scientific discovery does not proceed in a simple fashion where we have a single hypothesis that we would like to test.

A classical example is in the analysis of gene expression data. We measure the expression of tens of thousands of genes and we would like to know if any of them are associated with some phenotype (for example whether a person has a disease or not). Typically, the way this is done is that the scientist does tens of thousands of hypothesis tests, and then reports the associations that are significant, i.e. reports the tests where the null hypothesis was rejected.

This is very problematic:

Suppose we had 1000 individual hypotheses to test, and conducted 1000 hypothesis tests, and for each of them rejected the null when the $p$-value was less than $\alpha = 0.05$. How many times would you expect to falsely reject the null hypothesis?

The answer is we would expect to reject the null hypothesis 50 times if all 1000 null hypotheses were true. So we really cannot report all the discovered associations (rejections) as significant because we expect many false rejections.

The multiple testing problem is behind a lot of the "reproducibility crisis" of modern science. Many results that have been reported significant cannot be reproduced simply because they are false rejections. Too many false rejections come from doing multiple testing but not properly adjusting your tests to reflect the fact that many hypothesis tests are being done[1].

The basic question is how to we adjust our p-value cutoffs to account for the fact that multiple tests are being done.

### 14.3.1 The Family-Wise Error Rate

We first need to define what the error control we desire is. Recall, the Type I error controls the probability of falsely rejecting the null hypothesis. We have seen that in order to control the Type I error we can simply threshold the p-value, i.e rejecting the null if the p-value $\leq \alpha$ controls the Type I error at $\alpha$.

One possibility (and we will discuss a different one in the next lecture) is that when a scientist does multiple tests we care about controlling the probability that we falsely reject *any* null hypothesis. This is called the Family-Wise Error Rate (FWER).

The FWER is the probability of falsely rejecting the null hypothesis even once amongst the multiple tests.

The basic question is then: how do we control the FWER?

### 14.3.2 Sidak correction

Suppose we do $p$ hypothesis tests, and want to control the FWER at $\alpha$.

The Sidak correction says we reject any test if the p-value is smaller than:

$$\text{p-value} \leq 1 - (1-\alpha)^{1/p} = \alpha_t,$$

so we reject any test if its p-value is less than $\alpha_t$.

The main result is that: if the p-values are all *independent* then the FWER $\leq \alpha$.

**Proof:** Let $q$ be the number of true null hypotheses. The probability of falsely rejecting a fixed test when the null hypothesis is true is $\alpha_t$, so we correctly retain it with probability $1 - \alpha_t$.

Since the p-values are all independent the probability of falsely rejecting any null hypothesis is:

$$\text{FWER} = 1 - (1-\alpha_t)^q \leq 1 - (1-\alpha_t)^p = \alpha.$$

[1]Too many false rejections can also arise from tests that do not properly control the $\alpha$ level but this is usually easier to detect/fix.

### 14.3.3 Bonferroni correction

The main problem with the Sidak correction is that it requires the independence of p-values. This is unrealistic especially if you compute the test statistics for the different tests on the same set of data. The Bonferroni correction instead uses the union bound to avoid this assumption.

The Bonferroni correction says we reject any test if the p-value is smaller than:

$$\text{p-value} \leq \frac{\alpha}{p}.$$

The main result is that: The FWER $\leq \alpha$.

**Proof:** Assume that the null hypothesis is true for $i_1, ..., i_q$. In this case,

$$\text{FWER} = \mathbb{P}\left(\bigcup_{k=1}^{q} \text{reject } H_{0i_k}\right) \leq \sum_{k=1}^{q} \mathbb{P}\left(\text{reject } H_{0i_k}\right) \leq \sum_{i=1}^{p} \mathbb{P}\left(\text{reject } H_{0i}\right) \leq \sum_{i=1}^{p} \frac{\alpha}{p} = \alpha,$$

where the first inequality follows from the union bound.

### 14.3.4 p-value fishing (a.k.a p-hacking)

We have discussed why and how one must take care when dealing with multiple hypothesis tests and the correct interpretation of the individual p-values. The (incorrect and invalid) practice of reporting individual p-values as significant without taking account of multiplicity is called "p-value fishing".

There is another form of p-value fishing that involves only one hypothesis but is also commonly seen in practice. Suppose we have a hypothesis testing problem $H_0$ and $H_1$ about the density $f_\theta$. Suppose we can obtain unlimited number of samples from $f_\theta$. Let $X_1, ..., X_n, ...$ be the potentially infinite data sequence. At a particular sample size $n$, I can calculate the $p$-value for the data set $(X_1, ..., X_n)$, denoted as $p_n$. The p-value fishing here is to increase the sample size until $p_n$ is below a certain threshold (for example, 0.05) and claim rejection of the null.

One can show that when $H_0$ is true, for any $\alpha \in (0, 1)$, with probability one there exists a sample size $N$ (although random) such that $p_N \leq \alpha$. Therefore, the meaning of $p$-value is no longer valid if we adjust the inference problem after seeing the data! There are active areas of research called "adaptive inference" and "post-selection inference" that tackle these kind of issues.