Lecture Notes 17
36-705

Now we will take a look at some specific hypothesis testing problems. And we shall depart

1 Goodness-of-fit testing

Let $X_1, \ldots, X_n \sim P$. We want to test:

$$\begin{aligned} H_0 : &\quad P = P_0 \\ H_1 : &\quad P \neq P_0, \end{aligned}$$

for some fixed, known distribution P_0. As an example, suppose that P is multinomials on k categories. The null distribution just a vector of probabilities $(p_{01}, \ldots, p_{0k})$, with $p_{0i} \geq 0$, $\sum_i p_{0i} = 1$. We could use the LRT but here we introduce another popular test.

Given a sample $X_1, \ldots, X_n$ you can reduce it to a vector of counts $(Z_1, \ldots, Z_k)$ where Z_i is the number of times we observed the i-th category. Let

$$T(X_1, \ldots, X_n) = \sum_{i=1}^{k} \frac{(Z_i - np_{0i})^2 - np_{0i}}{np_{0i}}.$$

On your HW you will show that asymptotically this test statistic, under the null, has a χ^2_{k-1} distribution. This is called Pearson's χ^2 test. More generally, we can perform any goodness-of-fit test by reducing to a multinomial test by binning, i.e. you define a sufficiently find partition of the domain, this induces a multinomial p_0 under the null which you then test using Pearson's test.

2 Two-sample Testing

Another popular hypothesis testing problem is the following: you observe $X_1, \ldots, X_{n_1} \sim P$ and $Y_1, \ldots, Y_{n_2} \sim Q$, and want to test if:

$$\begin{aligned} H_0 : &\quad P = Q \\ H_1 : &\quad P \neq Q. \end{aligned}$$

Assume first that P and Q are in the same parametric family $(P_\theta : \ \theta \in \Theta)$ So $p = p(x; \theta_1)$ and $q = p(x; \theta_2)$ for some θ_1, θ_2. We want to test $H_0 : \theta_1 = \theta_2$. If the parameter is scalar, the Wald test statistic is

$$T = \frac{\widehat{\theta}_1 - \widehat{\theta}_2}{se}$$

where

$$se^2 = \frac{se_1^2}{n_1} + \frac{se_2^2}{n_2}$$

and se_1 and se_2 are estimates of the standard errors of $\widehat{\theta}_1$ and $\widehat{\theta}_2$. If θ is a vector, we can use the $LRT = -2\log\lambda$ where

$$\lambda = \frac{\prod_i p(X_i; \widehat{\theta}) \prod_i p(Y_i; \widehat{\theta})}{\prod_i p(X_i; \widehat{\theta}_1) \prod_i p(Y_i; \widehat{\theta}_2)}$$

where $\widehat{\theta}$ is the mle under H_0 obtained by combining both samples. If θ is a vector of length k, then under the null there are k parameters and under the alternative there are $2k$ parameters. The difference is k. So the LRT converges to χ^2_k under H_0.

Consider again the multinomial setting where P and Q are multinomials on k categories. Then there is a version of the χ^2 test that is commonly used. Let us define $(Z_1, \ldots, Z_k)$ and $(Z'_1, \ldots, Z'_k)$ to be the counts in the X and Y sample respectively. We can define for $i \in \{1, \ldots, k\}$,

$$\widehat{c}_i = \frac{Z_i + Z'_i}{n_1 + n_2}.$$

The two-sample χ^2 test is then:

$$T_n = \sum_{i=1}^{k} \left[\frac{(Z_i - n_1\widehat{c}_i)^2}{n_1\widehat{c}_i} + \frac{(Z'_i - n_2\widehat{c}_i)^2}{n_2\widehat{c}_i} \right].$$

This is a bit harder to see but under the null this statistic also has a χ^2_{k-1} distribution.

3 The Permutation Test

For two-sample testing we can determine the cutoff in a different way without resorting to asymptotics and without assuming a parametric model.

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.

Let $T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ be any test statistic. For example,

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

Let us denote the value of the test statistic computed on the observed data by T_{obs}.

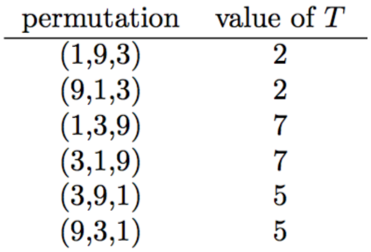
The idea of the permutation test is simple. Define $N = m + n$ and consider all $N!$ permutations of the data $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$. For each permutation we could compute our test statistic T. Denote these as $T_1, \ldots, T_{N!}$. The key observation is: **under the null hypothesis 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 α controls the Type I error at α. Here is a toy-example:

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

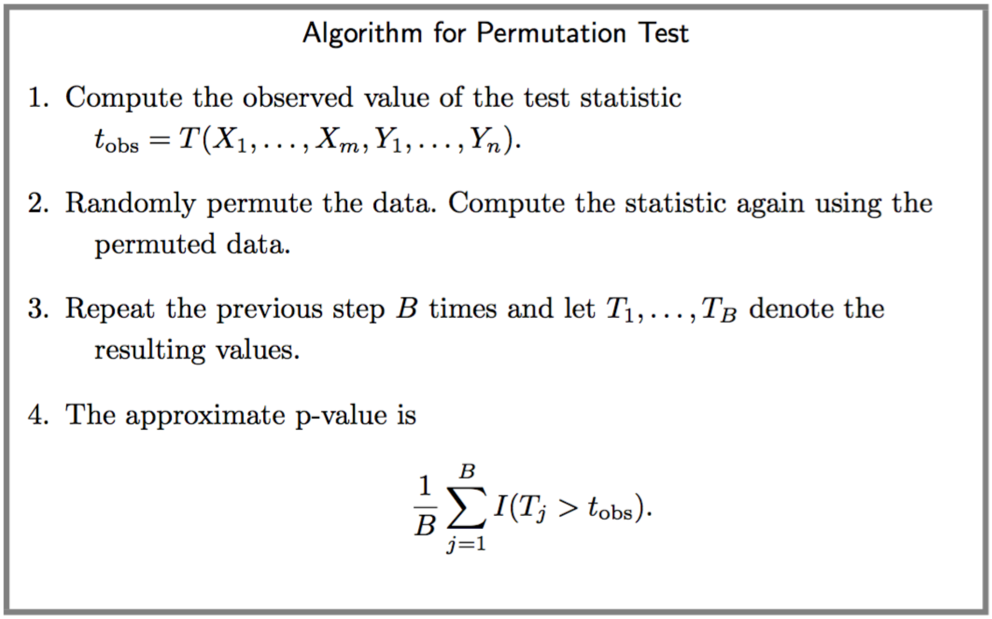


permutation	value of T
(1,9,3)	2
(9,1,3)	2
(1,3,9)	7
(3,1,9)	7
(3,9,1)	5
(9,3,1)	5

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

$$\text{p-value} = \frac{4}{6} = 0.66,$$

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 for large N. 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

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

We first show that the permutation test that we covered last time actually controls the Type I error, and then move on to the problem of multiple testing which will occupy us for a couple of lectures.

4 Analyzing the permutation test for two-sample testing

We observe $X_1, \ldots, X_{n_1} \sim P$ and $Y_1, \ldots, Y_{n_2} \sim Q$, and want to test if:

$$\begin{aligned} H_0 : &\quad P = Q \\ H_1 : &\quad P \neq Q. \end{aligned}$$

Let us introduce some notation: we suppose we are given a test statistic T which is a function of the observed data, for instance:

$$T(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) = \left| \frac{1}{n_1} \sum_{i=1}^{n_1} X_i - \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i \right| := t_{\text{obs}}.$$

We let $N = n_1 + n_2$, and denote the permutations of the data by $\{Z_1, \ldots, Z_{N!}\}$. We let:

$$\phi_{\text{perm}}(Z_{\text{obs}}) = \mathbb{I}\left[\left(\frac{1}{N!} \sum_{i=1}^{N!} \mathbb{I}(T(Z_i) > t_{\text{obs}}) \right) < \alpha \right].$$

We claim that:

$$\mathbb{P}_{H_0}(\phi_{\text{perm}}(Z_{\text{obs}}) = 1) \leq \alpha.$$

Proof: Note that the permutation test would reject the null only for test statistics that are in the upper α-quantile of the distribution of test statistics, i.e.:

$$\alpha \geq \frac{1}{N!} \sum_{i=1}^{N!} \phi_{\text{perm}}(Z_i).$$

Taking the expectation over Z_i under the null we obtain that,

$$\alpha \geq \frac{1}{N!} \sum_{i=1}^{N!} \mathbb{E}_{H_0}[\phi_{\text{perm}}(Z_i)].$$

Under the null hypothesis each dataset Z_i has the same distribution as Z_{obs} so we obtain that:

$$\alpha \geq \frac{1}{N!} \sum_{i=1}^{N!} \mathbb{E}_{H_0}[\phi_{\text{perm}}(Z_{\text{obs}})],$$

i.e. that,

$$\mathbb{P}_{H_0}(\phi_{\text{perm}}(Z_{\text{obs}}) = 1) \leq \alpha,$$

as desired. Also note that up to some small quantization error (since the p-values that the permutation test produces are discrete), all of the above inequalities are actually equalities, i.e. the permutation test has Type I error that is very close to α.

5 Multiple Testing

Testing many hypotheses at once, is called 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 good 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 did 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. So we really cannot report all the discovered associations (rejections) as significant because we expect many false rejections.

Another example is in vaccine trials. If we keep testing whether a vaccine is effective as time goes on, we will end up doing many tests.

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

5.1 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 α.

One possibility is to control 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?

5.2 Sidak correction

Suppose we do d hypothesis tests, and want to control the FWER at α. The Sidak correction says to reject any test if the p-value is smaller than:

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

so we reject any test if its p-value is less than α_t.

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

Proof: Suppose that all the null hypotheses are true (this is called the *global null*). You can easily see that if this is not the case you can simply ignore all the tests for which the null is false. The probability of falsely rejecting a fixed test is α_t, so we correctly fail to reject it with probability $1 - \alpha_t$.

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

$$\text{FWER} = 1 - (1 - \alpha_t)^d = \alpha.$$

5.3 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:

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

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

Proof: Suppose again that the global null is true. In this case,

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

where the first inequality follows from the union bound.

5.4 Holm's procedure

There are many possible improvements to the Bonferroni procedure. For instance, suppose that I told you that exactly (or at most) d_0 of the null hypotheses are truly nulls. Then you can see that we could have used the cut-off of $\frac{\alpha}{d_0}$ and still maintained control over the FWER.

As a thought experiment consider the following setting. You conduct $d = 5$ experiments and you observe p-values of $(0.7, 0.02, 0, 0, 0)$.

Intuitively, it seems like since we are absolutely sure that the last three experiments are non-nulls we should be able to use the cut-off of $\alpha/2$ for the remaining two tests, and still control the FWER.

At a high-level it seems intuitively clear to us that other p-values for $\{p_j\}_{j \neq i}$ contain information at least about the number of null hypotheses and we can use this to relax the correction for p_i. Holm's procedure translates this intuition into a rigorous procedure.

1. Order the p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$.
2. If $p_{(1)} < \frac{\alpha}{d}$ then reject $H_{(1)}$ and move on, else stop and accept all H_i.
3. If $p_{(2)} < \frac{\alpha}{d-1}$ then reject $H_{(2)}$ and move on, else stop and accept $H_{(2)}, \ldots, H_{(d)}$.

   $\vdots$

4. If $p_{(d)} < \alpha$, then reject $H_{(d)}$, else accept $H_{(d)}$.

More succinctly, let

$$i^* = \min \left\{ i : p_{(i)} > \frac{\alpha}{d-i+1} \right\},$$

and reject all $H_{(i)}$ for $i < i^*$.

Holm's procedure controls the FWER at level α. Importantly, Holm's procedure does not require independence of the p-values, and it strictly dominates the Bonferroni procedure.

Proof: Let I_0 denote the indices of the true nulls. First let us make an observation: if

$$\min_{i \in I_0} p_i > \frac{\alpha}{d_0},$$

then we reject none of the true nulls. This is because the first time we encounter a true null we would compare it to a threshold that is at most α/d_0, and if we fail to reject it we would not reject any of the other true nulls.

So the FWER is:

$$\text{FWER} \leq \mathbb{P}\left(\min_{i \in I_0} p_i \leq \frac{\alpha}{d_0} \right) \leq \alpha,$$

by the union bound.

5.5 Something to think about

In the above discussion we assumed that there was a single scientist doing a bunch of tests so he could appropriately correct his procedure for the multiple testing problem. One thing to ponder is really what error rate should we be controlling, i.e. maybe I am the editor of a journal, and I want to ensure that across all articles in my journal the FWER is $\leq \alpha$. Maybe I want this to be true across the entire field? Should I be adjusting my p-values for people in other disciplines? Sounds absurd but it actually makes sense if you think about each of these procedures and their implications for reproducibility.

6 False Discovery Rate

Suppose that we tested $d = 1000$ genes for association with some disease, we got a 1000 p-values, and 100 of them were less than 0.01. We'd expect that roughly $0.01d_0 \leq 0.01d = 10$ of these to be falsely rejected nulls, and perhaps this is not a bad tradeoff, i.e. if we rejected 100 nulls, we would spend only 10% of our time on falsely rejected nulls, i.e. we would make 90 real discoveries.

This suggests using a different error criterion. The FDR (false discovery rate) is the expected number of false rejections divided by the number of rejections.

Denote the number of false rejections as V, and the total number of rejections as R. Then the false discovery *proportion* is:

$$\text{FDP} = \begin{cases} \frac{V}{R} & \text{if} \quad R > 0 \\ 0 & \text{if} \quad R = 0. \end{cases}$$

The FDR is then defined as:

$$\text{FDR} = \mathbb{E}[\text{FDP}].$$

In this notation we can see that the FWER is:

$$\text{FWER} = \mathbb{P}(V \geq 1).$$

We will next consider how one can control the FDR. We will describe a procedure known as the Benjamini-Hochberg (BH) procedure.