36-705: Intermediate Statistics **Fall 2019**

Lecture 22: October 21

Lecturer: Siva Balakrishnan

We begin discussing p-values, and then move on to discussing some important special testing problems.

22.1 p-values

When we test at a given level α we will reject or not reject. It is useful to summarize what levels we would reject at and what levels we would not reject at.

The p-value is the smallest α at which we would reject H_0.

In other words, we reject at all $\alpha \geq p$. So, if the pvalue is 0.03, then we would reject at $\alpha = 0.05$ but not at $\alpha = 0.01$.

Hence, to test at level α, we reject when $p < \alpha$.

Theorem 22.1 *Suppose we have a test of the form: reject when $T(X_1, \ldots, X_n) > c$. Then the p-value is*

$$p = \sup_{\theta \in \Theta_0} P_\theta(T_n(X_1, \ldots, X_n) \geq T_n(x_1, \ldots, x_n))$$

where $x_1, \ldots, x_n$ are the observed data and $X_1, \ldots, X_n \sim p_\theta$.

Example 22.2 *$X_1, \ldots, X_n \sim N(\theta, 1)$. Test that $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We reject when $|T_n|$ is large, where $T_n = \sqrt{n}(\overline{X}_n - \theta_0)$. Let t_n be the obsrved value of T_n. Let $Z \sim N(0, 1)$. Then,*

$$p = P_{\theta_0}\left(|\sqrt{n}(\overline{X}_n - \theta_0)| > t_n\right) = P(|Z| > t_n) = 2\Phi(-|t_n|).$$

The p-value is a random variable. Under some assumptions that you will see in your HW the p-value will be uniformly distributed on $[0, 1]$ under the null.

Important. Note that p is NOT equal to $\mathbb{P}(H_0|X_1, \ldots, X_n)$. The latter is a Bayesian quantity which we will discuss later.

22.2 More on p-values

In this section we will consider a simple example, derive the p-value, and derive its distribution under the null (you will do some of this more generally in your HW).

Suppose that $X_1, \ldots, X_n \sim N(\theta, 1)$ and we want to distinguish:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0.$$

We consider the test statistic: $T_n = \sqrt{n}(\overline{X}_n - \theta_0)$, and notice that under the null $T_n \sim N(0, 1)$. We reject for large values of T_n, and we can verify that the p-value is given by:

$$\text{p-value} = \Phi(-T_n).$$

Notice that the p-value is a random variable which takes values in $[0, 1]$. Now, we can derive its distribution *under the null*:

$$\mathbb{P}_0(\text{p-value} \leq u) = \mathbb{P}_0(\Phi(-T_n) \leq u) = \mathbb{P}_0(-T_n \leq \Phi^{-1}(u)) = \Phi(\Phi^{-1}(u)) = u,$$

where we have use the fact that Φ is continuous and increasing, and that under the null $-T_n$ has a $N(0, 1)$ distribution.

22.3 Goodness-of-fit testing

There are many testing problems that go beyond the usual parameteric testing problems we have formulated so far. Since we may not get a chance to cover these in detail later on, I thought it might be useful to discuss them briefly here.

The most canonical non-parametric testing problem is called goodness-of-fit testing. Here given samples $X_1, \ldots, X_n \sim P$, we want to test:

$$H_0 : \quad P = P_0$$
$$H_1 : \quad P \neq P_0,$$

for some fixed, known distribution P_0.

As a hypothetical example, you collect some measurements from a light source, you believe that the number of particles per unit time should have a Poisson distribution with a certain rate parameter (the intensity), and want to test this hypothesis.

22.3.1 The χ^2 test

In the simplest setting, P_0 and P are multinomials on k categories, i.e. the null distribution just a vector of probabilities $(p_{01}, \ldots, p_{0k})$, with $p_{0i} \geq 0$, $\sum_i p_{0i} = 1$.

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 you observed the i-th category.

A natural test statistic in this case (you could also do the likelihood ratio test) is to consider:

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

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, you could do 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.

22.4 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:

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

There are many popular ways of testing this (for instance, in the ML literature kernel-based tests are quite popular – search for "Maximum Mean Discrepancy" if you are curious).

Suppose again we considered 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. For two-sample testing we can determine the cutoff in a different way without resorting to asymptotics. This is called a permutation test. We will explore this in the general (not multinomial) setting.

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_{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* α 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 α controls the Type I error at α. We will return to a formal proof of this fact in a subsequent lecture.

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 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

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