

**Economics 696F, Causal Inference and Program Evaluation**  
**Lecture Note 5: Nonparametric Testing Notes**

**References:**

Conover, W. J., *Practical Nonparametric Statistics*, 3rd ed., Wiley.

Hajek, J., Sidak, Z., and Sen, P. K., 1999, *Theory of Rank Tests*, Academic Press.

Van der Vaart, A. W., 1998, *Asymptotic Statistics*, Cambridge University Press.

**Heuristics for Constructing a Test**

1. What possible probability distributions could describe the data? (Might be a parametric family, or a much larger set of distributions).
2. Divide the set of possible distributions into two subsets: the null hypothesis set and the alternative set.
3. Construct a test statistic – a function of the test statistic. Since the test statistic is a function of the random data, it will have a distribution that depends on the distribution of the data. We want the test statistic to have some nice properties:
  - For every distribution of the data under the null, we should be able to calculate the distribution of the test statistic fairly easily.
  - It's especially nice if the distribution of the test statistic is the same for every data distribution in the null.
  - If the distribution of the data falls in the alternative set, the test statistic should have a distribution that somehow differs from the distributions under the null.

**Formal Setup:**

We observe a random vector

$$Z \sim F_\theta, \quad \theta \in \Theta,$$

where  $\theta$  is a parameter. So we can think of  $\Theta$  as the set of possible probability distributions for  $Z$ .

We should think of  $Z$  as the entire data set. For example, we might observe  $n$  observations on some outcome:  $Y_1, \dots, Y_n$ . So let

$$Z = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

Suppose we assume

$$Y_i \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2),$$

then  $\theta = (\mu, \sigma^2)$ ,  $\Theta = \mathbb{R} \times \mathbb{R}_+$ , and

$$Z \sim \left( \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}, \sigma^2 I_n \right).$$

Hypotheses:

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \subset \Theta, \\ H_1 &: \theta \in \Theta_0^c. \end{aligned}$$

Test Statistic: a function of data  $t(Z)$ .

Critical Region: reject  $H_0$  if  $t(Z) \in Cr$ .

Significance Level:  $\alpha$  (e.g.  $\alpha = 0.05$ ).

We require: for **all**  $\theta \in \Theta_0$ ,

$$P_\theta(t(Z) \in Cr) \leq \alpha.$$

LHS is “Type I Error” probability.

Power Function:

$$\beta(\theta) = P_\theta(t(Z) \in Cr), \quad \theta \in \Theta_0^c.$$

Note power =  $1 - P_\theta(\text{Type II error})$ .

In general, we want high power (low Type II error probability) subject to constraint that Type I error probability is less than or equal to  $\alpha$ .

So to construct a test, we look for a statistic where:

- We can calculate its distribution—exact or asymptotic—under *all*  $\theta \in \Theta_0$ . That way we can calculate Type I error probabilities, and choose  $Cr$  appropriately.
- The distribution is sensitive to deviations from  $H_0$ , so that power is nontrivial.

## Two Sample Problem

Two treatments (0, 1).

Observe some outcome  $Y_i$ .

Order the observations:

- $Y_1, \dots, Y_{n_0}$  receive treatment 0.

- $Y_{n_0+1}, \dots, Y_{n_0+n_1}$  receive treatment 1.

Let  $n = n_0 + n_1$ .

To put in our framework, we need to specify  $\Theta$ .

### Parametric Example

Suppose we assume:

$$Y_1, \dots, Y_{n_0} \stackrel{\text{i.i.d.}}{\sim} N(\mu_0, \sigma_0^2),$$

$$Y_{n_0+1}, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} N(\mu_1, \sigma_1^2).$$

Assume independence between the two groups of observations.

Also assume (just for simplicity) that  $\sigma_0^2$  and  $\sigma_1^2$  are known. So:

$$Z = (Y_1, \dots, Y_n)',$$

$$\theta = (\mu_0, \mu_1), \quad \Theta = \mathbb{R}^2.$$

Hypotheses:

$$H_0 : \quad \mu_0 = \mu_1$$

$$H_1 : \quad \mu_0 \neq \mu_1.$$

So  $H_0$  is the line of all points  $(\mu_0, \mu_1)$  s.t.  $\mu_0 = \mu_1$ .

Start with

$$\bar{Y}_0 := \frac{1}{n_0} \sum_{i=1}^{n_0} Y_i, \quad \bar{Y}_1 := \frac{1}{n_1} \sum_{i=n_0+1}^n Y_i.$$

Under our assumptions,

$$\bar{Y}_0 \sim N\left(\mu_0, \frac{\sigma_0^2}{n_0}\right), \quad \bar{Y}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right),$$

and  $\bar{Y}_0$  and  $\bar{Y}_1$  are independent. So

$$(\bar{Y}_0 - \bar{Y}_1) \sim N\left(\mu_0 - \mu_1, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right),$$

and

$$\frac{(\bar{Y}_0 - \bar{Y}_1)}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} \sim N(\mu_0 - \mu_1, 1).$$

Let

$$t(Z) := \frac{(\bar{Y}_0 - \bar{Y}_1)}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}}.$$

Under  $H_0$ :

$$t(Z) \sim N(0, 1).$$

Important: distribution of  $t(Z)$  is the same for *all* parameter values in the null hypothesis. We say that  $t(Z)$  is *pivotal*.

So could use

$$Cr = (-\infty, -1.96] \cup [1.96, \infty).$$

For all  $\mu_0, \mu_1$  such that  $\mu_0 = \mu_1$ ,

$$P_{\mu_0, \mu_1}(t(Z) \in Cr) = 0.05.$$

### One-sided Test

Suppose we want to instead test:

$$H_0 : \mu_0 \leq \mu_1,$$

vs.

$$H_1 : \mu_0 > \mu_1.$$

We can use the same test statistic  $t(Z)$ . Now, however, its distribution will depend on the specific distribution in the null. For example, if  $\mu_0 - \mu_1 = -1$ , then  $t(Z)$  will have mean -1.

Suppose we use a critical region

$$Cr = [1.96, \infty).$$

Then, if  $\mu_0 - \mu_1 = 0$ ,

$$Pr(t(Z) \in Cr) = 0.025.$$

If instead, say,  $\mu_0 - \mu_1 = -1$ , then

$$Pr(t(Z) \in Cr) < 0.025.$$

So for all  $\mu_0 - \mu_1 \leq 0$ ,

$$Pr(t(Z) \in Cr) \leq 0.025.$$

So the test has significance level 0.025.

### Nonparametric Model and Asymptotic Test

Suppose we don't want to make the assumption that the  $Y_i$  are normally distributed.

Let  $\mathcal{F}$  be the set of all univariate distributions with finite mean and variance.

Suppose

$$Y_1, \dots, Y_{n_0} \stackrel{\text{i.i.d.}}{\sim} F_0,$$
$$Y_{n_0+1}, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} F_1,$$

where  $F_0, F_1 \in \mathcal{F}$ . Also assume that all the observations in treatment group 0 are independent of all the observations in treatment 1.

We can think of  $\Theta$  as the set of all possible pairs of distributions with finite mean and variance

$$\Theta = \mathcal{F} \times \mathcal{F}.$$

This is a very large “parameter” space!

Hypotheses:

$$H_0 : \mu(F_0) = \mu(F_1),$$

where  $\mu(F_0)$  stands for the expected value under  $F_0$ , etc.

$$H_1 : \mu(F_0) \neq \mu(F_1).$$

If we know the variance of  $F_0$  and  $F_1$ , we might consider using our test statistic  $t(Z)$ . The problem is that the distribution of  $t(Z)$  depends on  $F_0$  and  $F_1$ . Even if  $\mu(F_0) = \mu(F_1)$ , it is *not* in general  $N(0, 1)$ .

So  $t(Z)$  is no longer pivotal, and it is very difficult to know its distribution under all possible distributions consistent with the null hypothesis.

One possibility is to try to use large-sample approximations.

Let

$$\hat{\sigma}_0^2 = \frac{1}{n_0} \sum_{i=1}^{n_0} (Y_i - \bar{Y}_0)^2,$$

$$\hat{\sigma}_1^2 = \frac{1}{n_1} \sum_{i=n_0+1}^n (Y_i - \bar{Y}_1)^2.$$

Suppose  $n_0 \rightarrow \infty$ ,  $n_1 \rightarrow \infty$ , and  $n_0/n \rightarrow \kappa \in (0, 1)$ .

Then

$$\begin{aligned} \bar{Y}_0 &\xrightarrow{p} \mu(F_0), \\ \bar{Y}_1 &\xrightarrow{p} \mu(F_1), \\ \hat{\sigma}_0^2 &\xrightarrow{p} \sigma^2(F_0), \\ \hat{\sigma}_1^2 &\xrightarrow{p} \sigma^2(F_1). \end{aligned}$$

Let

$$\tilde{t}(Z) := \frac{(\bar{Y}_0 - \bar{Y}_1)}{\sqrt{\frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1}}}.$$

Can show that, for large  $n_0, n_1$ , the distribution of  $\tilde{t}(Z)$  is well approximated by  $N(0, 1)$ , for all distributions in the null hypothesis. (We say that  $\tilde{t}(Z)$  is asymptotically pivotal.)

So we can use the same critical region as before, and have a test that will be approximately valid.

Note: we could view this as a test of the hypotheses:

$$H_0 : F_0 = F_1,$$

vs

$$H_1 : F_0 \neq F_1.$$

The (asymptotic) distribution under the null hypothesis will be the same. Now, however, there may be pairs of distributions in  $H_1$  such that their expected values are equal, so that the test would have low power against those alternatives.

### Rank Tests

If  $n_0, n_1$  are not large, it would be nice to have a test that does not depend on asymptotic approximations for their validity.

Given  $Y_1, \dots, Y_n$ , let  $R_1, \dots, R_n$  be the ranks of the values. For example, if  $Y_4$  is the lowest among all the  $Y_i$ 's, then  $R_4 = 1$ .

If two (or more) observations are tied, make their rank equal to the average of the tied ranks.

Assume, as before, that

$$\begin{aligned} Y_1, \dots, Y_{n_0} &\stackrel{\text{i.i.d.}}{\sim} F_0, \\ Y_{n_0+1}, \dots, Y_n &\stackrel{\text{i.i.d.}}{\sim} F_1, \end{aligned}$$

where  $F_0, F_1 \in \mathcal{F}$ , and assume independence between the two groups of observations.

Let the null hypothesis be:

$$H_0 : F_0 = F_1.$$

The Wilcoxon/Mann Whitney test statistic is:

$$t(Z) = \sum_{i=1}^{n_0} R_i.$$

This is the sum of the ranks of only the group that got treatment 0. Intuitively, if the two distributions  $F_0$  and  $F_1$  are the same, then

$$Y_1, \dots, Y_n \sim F = F_0 = F_1,$$

and the ranks of the treatment 0 group should not be systematically higher or lower than the ranks of treatment 1 group. So if we see an unusually high or low value of  $t(Z)$ , we would interpret this as evidence inconsistent with the null hypothesis.

Amazingly, if  $F_0 = F_1 = F$ , and the distribution is continuous, then the distribution of  $t(Z)$ :

- Is the same for any continuous  $F$ ,
- can be easily tabulated or simulated on the computer.

So we have a “nonparametric” test which does not require asymptotic approximations to be valid.

The key result underlying the distribution of the Wilcoxon test:

Lemma: *Suppose that  $Y_1, \dots, Y_n$  are IID from a continuous distribution  $F$ . Let  $R = (R_1, \dots, R_n)$  be the ranks of  $Y_1, \dots, Y_n$ . Then the distribution of  $R$  puts equal probability on each of the  $n!$  permutations of  $(1, 2, \dots, n)$ .*

For a proof of this result, see Hajek, Sidak, and Sen.

If we wanted to simulate the distribution of  $t(Z)$  under the null hypothesis, we could:

1. Randomly generate a permutation of  $(1, 2, \dots, n)$ .
2. Sum the first  $n_0$  numbers in the permutation to get a draw for  $t(Z)$ .
3. Repeat many times to get an approximate distribution for  $t(Z)$  under the null.

In practice we do not need to do this, because the distribution has already been tabulated and there exist standard tables for the Wilcoxon/Mann-Whitney test. However, these are generally valid only if the distribution is continuous (meaning in practice, no ties in the ranks.)

What if  $F$  is not continuous? If there are only a few ties, the standard tables are not exactly correct, but give good approximations in practice.

If there are many ties, the standard tables are not valid. It would be possible in principle to work out the exact distribution of  $t(Z)$  under the null, but impractical to have separate tables for every possibility. One could simulate the p-values, or turn to asymptotic approximations. It turns out that the Wilcoxon test also has an asymptotically normal distribution, under weak conditions; see van der Vaart (1998), Ch. 13. The asymptotic normal approximation actually works quite well even for moderate sample sizes.

The Wilcoxon/Mann-Whitney test can be extended to testing if the distribution in more than 2 groups is equal. See Conover, 5.2.