IEOR 151 – Lecture 4
Location Tests

1 Varieties of Tests

There are a large number of hypothesis tests designed to compare the means (or median or mode) of (i) one group to a constant, (ii) of two groups to each other, or (iii) of multiple groups to each other. Even for the common setup of comparing the mean of two groups, tests can be classified as those that are paired or unpaired and those in which the variances of the two groups are equal or unequal under the null hypothesis. In a paired setup, data from the two groups are matched in a one-to-one manner; an example is measuring blood pressure from a patient before and after they start taking a medication. The large number of different tests can be confusing, but there is an important reason for their distinctions. Depending on the distributions of the data assumed by the null hypothesis and the number of data points, certain tests will have greater power than others.

To better appreciate these differences, we will discuss location tests in which the means of two groups are compared to each other and in which the data is unpaired. Two important tests for this situation are the t-test and the Mann-Whitney U test. For simplicity, we will focus on the situation in which the variances of the two groups are equal under the null hypothesis. More specifically, we are interested in the "base" null hypothesis

$$h_0 : \mathbb{E}(A) = \mathbb{E}(B) \text{ assuming that } \mathrm{Var}(A) = \mathrm{Var}(B),$$

for independent measurements from two groups with distributions $A_i \sim F_a(u)$ and $B_i \sim F_b(u)$. Furthermore, we will assume that we have n_1 measurements from group A and n_2 measurements from group B.

2 Unpaired Two-Sample t-Test

The reason that there are so many hypothesis tests that are called t-tests is that this name is given to any test whose test statistic follows the Student's t distribution. Confusingly, this name is also given to tests whose test statistic approximately follows the Student's t distribution; an example of this is Welch's t-test, which is used for the case where the variances of the two groups are assumed to be unequal. Here, we will focus on two cases.

But before we consider those two cases, it is useful to give some definitions. The *density* function of a Student's t distribution with $\nu > 0$ degrees of freedom is given by

$$f(u) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where $\Gamma(\cdot)$ is the gamma function. An alternative characterization is that the Student's t distribution is the distribution of the random variable T defined as

$$T = \frac{Z}{\sqrt{V/\nu}},$$

where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_\nu$ has a chi-squared distribution with ν degrees of freedom.

2.1 Normality of A and B

In the first case, we will amend the base null hypothesis with normality. That is, our null hypothesis for this section will be

$$H_0 : h_0 \text{ and } A_i, B_i \sim \mathcal{N}(\mu, \sigma^2),$$

where $\mu = \mathbb{E}(A) = \mathbb{E}(B)$ and $\sigma^2 = \text{Var}(A) = \text{Var}(B)$. A test statistic for this null hypothesis is given by

$$t = \frac{(\overline{A}_{n_1} - \overline{B}_{n_2})/\sqrt{1/n_1 + 1/n_2}}{\sqrt{S_{ab}/(n_1+n_2-2)}},$$

where we define

$$S_{ab} = \sqrt{(n_1-1)S_a^2 + (n_2-1)S_b^2)},$$

with $S_a^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(A_i - \overline{A}_{n_1})^2$ and $S_b^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(B_i - \overline{B}_{n_1})^2$. We will give some intuition for why the statistic t has a Student's t distribution with n_1+n_2-2 degrees of freedom. The original derivation by Fisher (though the test was invented by Gosset working at the Guinness brewery in Dublin, Ireland) uses the rotational symmetry of the problem to essentially provide a geometric argument. We will take an alternative approach:

1. Using results about jointly Gaussian random variables we can show that $\overline{A}_{n_1}$ and $A_i - \overline{A}_{n_1}$ (as well as $\overline{B}_{n_1}$ and $B_j - \overline{B}_{n_1}$) are independent. Since the numerator is a function of $\overline{A}_{n_1}$ and $\overline{B}_{n_1}$, and the denominator is a function of $A_j - \overline{A}_{n_1}$ and $B_j - \overline{B}_{n_1}$, the numerator and denominator are independent.

2. It can be shown that $S_a^2 \sim \sigma^2/(n_1-1)\cdot\chi^2_{n_1-1}$ and $S_b^2 \sim \sigma^2/(n_2-1)\cdot\chi^2_{n_2-1}$. But because $\chi^2_{k_1} + \chi^2_{k_2} \sim \chi^2_{k_1+k_2}$ for independent $\chi^2_{k_1}$ and $\chi^2_{k_2}$, we have that $(n_1-1)S_a^2 + (n_2-1)S_b^2 \sim \sigma^2\chi^2_{n_1+n_2-2}$.

3. Note that $\overline{A}_{n_1} - \overline{B}_{n_2} \sim \mathcal{N}(0, \sigma^2/n_1 + \sigma^2/n_2) = \mathcal{N}(0, \sigma^2 \cdot (1/n_1 + 1/n_2))$.
4. Last, observe that $t \sim \sigma\mathcal{N}(0,1)/\sqrt{\sigma^2\chi^2_{n_1+n_2-2}/(n_1+n_2-2)}$. But using the alternative characterization, this means that t is described by the Student's t distribution with n_1+n_2-2 degrees of freedom.

2.2 Non-Normal A and B with Large Sample Sizes

In the second case, we will amend the base null hypothesis with non-normality. That is, our null hypothesis for this section will be

$$H_0 : h_0 \text{ and } A_i, B_i \text{ are non-normal but have finite variance } \sigma < \infty\ .$$

It is the case that for large sample sizes, the test statistic t defined above is well approximated by a normal distribution $\mathcal{N}(0,1)$. To see why, suppose that $n_2 = \alpha n_1$ for some $\alpha \in (0,1)$. Next, note the following:

1. The Central Limit Theorem implies that $\sqrt{n_1}(\overline{A}_{n_1} - \mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$ and $\sqrt{n_2}(\overline{B}_{n_2} - \mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$. Next note that
$$\begin{aligned}(\overline{A}_{n_1} - \overline{B}_{n_2})/\sqrt{1/n_1+1/n_2} &= (\overline{A}_{n_1} - \mu + \mu - \overline{B}_{n_2})/\sqrt{1/n_1+1/n_2} \\ &= \frac{\sqrt{n_1}}{\sqrt{1+n_1/n_2}}(\overline{A}_{n_1} - \mu) - \frac{\sqrt{n_2}}{\sqrt{n_2/n_1+1}}(\overline{B}_{n_2} - \mu).\end{aligned}$$
But because (i) $1/\sqrt{1+n_1/n_2} = 1/\sqrt{1+1/\alpha}$, (ii) $1/\sqrt{1+n_2/n_1} = 1/\sqrt{1+\alpha}$, and (iii) A_i, B_i are independent, we have from the Continuous Mapping Theorem that
$$(\overline{A}_{n_1} - \overline{B}_{n_2})/\sqrt{1/n_1+1/n_2} \xrightarrow{d} \mathcal{N}(0, \sigma^2/(1+1/\alpha) + \sigma^2/(1+\alpha)) = \mathcal{N}(0,\sigma^2),$$
since some algebra gives that $1/(1+1/\alpha) + 1/(1+\alpha) = 1$.
2. Next, note that some algebra gives that $\mathbb{E}(S_a^2) = \sigma^2$. Also, note that we can rewrite S_a^2 as
$$\begin{aligned}S_a^2 &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(A_i - \overline{A}_{n_1})^2 \\ &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(A_i^2 - 2A_i\overline{A}_{n_1} + \overline{A}_{n_1}^2) \\ &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}A_i^2 - \frac{2\overline{A}_{n_1}}{n_1-1}\sum_{i=1}^{n_1}A_i + \frac{n_1}{n_1-1}\overline{A}_{n_1}^2.\end{aligned}$$
Now by the weak law of large numbers, we have that $1/n_1\sum_{i=1}^{n_1}A_i^2 \xrightarrow{p} \mathbb{E}(A_i^2)$ and $\overline{A}_{n_1} \to \mu$. Thus, using the Continuous Mapping Theorem gives
$$S_a^2 \xrightarrow{p} \mathbb{E}(A_i^2) - 2\mu^2 + \mu^2 = \mathbb{E}(A_i^2) - \mu^2 = \sigma^2,$$
where the equivalence to σ^2 is by definition of variance.

3. Finally, we note that some algebra yields $\sqrt{S_{ab}/(n_1+n_2-2)} \xrightarrow{p} \sigma$. Using Slutsky's Theorem to put everything together, we get that $t \xrightarrow{d} \mathcal{N}(0,1)$.

3 Mann-Whitney U Test

For this hypothesis test, we will add additional elements to the hypothesis test. One situation in which the Mann-Whitney U Test can be applied is if the null hypothesis is

$$H_0 : h_0 \text{ and } F_a(u) = F_b(u).$$

In fact, this test is not applicable if the distributions have different variances but the same means.

The test works as follows. The data A_i, B_i is placed into a single list that is sorted by ascending order, and the rank of a data point is defined as its ordinal position in the single list. Next, the sum of the ranks for all data points A_i is computed – call this value R_a. The test statistic U is defined as

$$U = \min\{R_a - n_1(n_1+1)/2, n_1 n_2 - R_a + n_1(n_1+1)/2\}.$$

The p-value is then computed using the distribution of U. The computation of this is beyond the scope of this class, but there are two points of intuition. The first is that the distribution of U can be approximated by $\mathcal{N}(n_1 n_2/2, n_1 n_2(n_1+n_2+1)/12)$. The second is that U is a measure of how evenly mixed the data is, with more extreme values indicating non-even mixing. If A_i, B_i have identical distributions, then you would expect the data to be evenly mixed.