IEOR 151 – Lecture 6
Multiple Comparisons

1 Example: Comparing Service Rates

Consider a situation in which there are four healthcare providers performing triage for an emergency room in a hospital. Triage is the process of evaluating the severity of a patient's condition and then assigning a priority for treatment. Each provider works one at a time. This is an essential element of emergency rooms because some patients will have a relatively benign condition such as a cold whereas other patients may be suffering from something more urgent like a heart attack.

It is common for a hospital to have a standardized procedure for triage in order to improve service rates and quality. Now suppose that three of the healthcare providers feel that the standardized procedure is suboptimal. As a result, these three have each made individual adjustments to the standardized triage procedure. There is concern that these deviations are resulting in higher mortality rates. To check this, mortality rates for each healthcare provider over multiple dates was collected.

A salient question to ask is what testing procedure to use. For instance, the average mortality rates for each pair of healthcare providers could be compared. This would entail a total of six comparisons of the average means. However, actually performing multiple tests is non-optimal because there is a loss of power associated with doing more than one test. What would be better is a method to *simultaneously* compare the four means. This would be better because it would comprise a single test, and so there would be no reduction in power due to having to perform multiple tests.

2 Analysis of Variance (ANOVA)

The idea of ANOVA is to simultaneously compare the mean (or median or mode) of different groups, under an assumption that the distributions of measurements from each group are identical. The situation is analogous to location tests in which the means of two groups are compared to each other. Under assumptions on the normality (i.e., Guassian or not) of the underlying distributions, different tests are available.

2.1 F-Test

Similar to the case with the t-test, an F-test is any test in which the test statistic follows an F-distribution. If $d_1, d_2 > 0$ are the degrees of freedom, then the corresponding density of the

F-distribution is given by

$$f(u) = \frac{\sqrt{\frac{(d_1 x)^{d_1} \cdot d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B(\frac{d_1}{2}, \frac{d_2}{2})},$$

where $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$ is the beta function. Again in analogy to the t-distribution, there is a simpler characterization of the F-distribution. If $U_1 \sim \chi^2_{d_1}$ and $U_2 \sim \chi^2_{d_2}$ are independent, then the random variable

$$X = \frac{U_1/d_1}{U_2/d_2}$$

has an F-distribution with $d_1, d_2 > 0$ degrees of freedom.

In the present case, the null hypothesis that we are interested in testing is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \text{ and } \sigma_1^2 = \sigma_2^2 = \ldots \sigma_k^2, \text{ where } X^i \sim \mathcal{N}(\mu_i, \sigma_i^2) \text{ for } i = 1, \ldots, k$$

Let n_i be the number of measurements taken from the i-th group, define $\overline{X}^i = 1/n_i \sum_{j=1}^{n_i} X^i_j$ and $\overline{X} = 1/(\sum_i n_i) \cdot \sum_{i=1}^k \sum_{j=1}^{n_i} X^i_j$. Then, the test statistic we use is $F = \frac{MSG}{MSE}$, where

$$MSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\overline{X}^i - \overline{X})^2$$

is the mean square between groups and

$$MSE = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X^i_j - \overline{X}^i)^2$$

is the mean squared error. The intuition is that MSG quantifies the amount of variation between groups, whereas MSE quantifies the amount of variation within groups. Under the null hypothesis, we would expect that the variation between and within groups should be equal. So in essence, we compute the p-value by looking for how far the test statistic F deviates from the value (of approximately) one, and this is computed using the F-distribution. Note that there is no notion of one-sided or two-sided here; the test statistic can only be positive and there is only one direction to test.

Just as was done for the t-test, some work shows that MSG is independent of MSE. More work shows that $MSG/\sigma^2 \sim \chi^2_{k-1}/(k-1)$ and $MSE/\sigma^2 \sim \chi^2_{n-k}/(n-k)$. Thus, it must be that F is described by an F-distribution with $d_1 = k-1$ and $d_2 = n-k$ degrees of freedom.

2.2 Kruskal–Wallis Test

Recall that the Mann-Whitney U test is a nonparametric hypothesis test for comparing two groups when their distributions are not Gaussian. The Kruskal–Wallis test is the extension of the Mann-Whitney U test to the situation with more than two groups. Here, the null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \text{ and } \sigma_1^2 = \sigma_2^2 = \ldots \sigma_k^2 \text{and} f_{X^i}(u) = f_{X^j}(u) \text{ for } i \neq j.$$

For simplicity, we will assume that no measured value occurs more than once in the data. The test works as follows. First, the data from every group is placed into a single list that is sorted into ascending order, and a rank from 1 to $N = \sum_{i=1}^{k} n_1$ is assigned to each data point in the single list. Next, the test statistic

$$K = (N-1)\frac{\sum_{i=1}^{k} n_i(\overline{r}^i - \overline{r})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(r_j^i - \overline{r})^2},$$

where r_j^i is the rank of the j-th data point from the i-th group, $\overline{r}^i = 1/n_i \cdot \sum_{j=1}^{n_i}$, and $\overline{r} = 1/N \cdot \sum_{i=1}^{k}\sum_{j=1}^{n_i} r_j^i = (N+1)/2$. The intuition is that this looks similar to the test statistic for the F-test, but here we are looking at a quantity that looks like the variation in rank between groups divided by the variation in rank within groups. A lot of algebra shows that

$$K = \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i(\overline{r}^i)^2 - 3(N+1),$$

which approximately looks like a χ^2_{k-1} distribution when the n_i are large.

3 Multiple Testing with Multiple Comparisons

Suppose that a multiple comparison is performed, and the null hypothesis that each group is identical is rejected. It is natural to ask which of the pairs of groups are different, but this necessitates comparing all pairs of groups. And doing so introduces a multiple testing situation. There are many ways to do corrections for the multiple tests, but one way is to compute p_{ij}-values for each pairwise comparison (say $k(k-1)/2$ pairwise tests) and a p-value for the ANOVA test. If $p < \alpha/2$, then the null hypothesis that all the groups are identical is rejected. And then, a multiple testing procedure (e.g., the Bonferroni correction or the Holm-Bonferroni method) to ensure the familywise error rate of the pairwise comparisions is below $\alpha/2$. Note that this ensures that the entire procedure ensures the familywise error rate is below α, because

$$\begin{aligned} FWER &= \mathbb{P}(p < \alpha/2 \cup \text{pairwise errors } < \alpha/2) \\ &\leq \mathbb{P}(p < \alpha/2) + \mathbb{P}(\text{pairwise errors } < \alpha/2) = \alpha/2 + \alpha/2. \end{aligned}$$

Note that we could have an infinite number of procedures by varying the condition to $p < \gamma\alpha$ and $FWER_{\text{pairwise}} < (1-\gamma)\alpha$ for $\gamma \in (0,1)$; however, the value γ should not depend upon the value p otherwise the derivation above may not hold. In other words, γ needs to be selected prior to conducting the tests.