# IEOR 165 – Lecture 20
# 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 were 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 the use of multiple testing adjustments (like the Bonferroni correction of the Holm-Bonferroni method) are conservative approaches to controlling the familywise error rate; they are conservative in the sense that the adjusted $p$-value will be larger than the true $p$-value would be if we were able to exactly take into account the multiple tests. 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 conservativeness introduced by performing a correction for multiple tests.

## 2 One-Way Analysis of Variance (ANOVA)

The idea of ANOVA is to simultaneously compare the mean (or median) of different groups, under an assumption that the distributions of measurements from each group are identical. The situation is analogous to two-sample location tests in which the means (or medians) of two groups are compared to each other. Depending on the distributional assumptions in the null hypothesis, 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. Recall that the notation $U \sim \chi^2(d)$ indicates that $U$ has a $\chi^2$-distribution with $d$ degrees of freedom, and let $U_1 \sim \chi^2(d_1)$ and $U_2 \sim \chi^2(d_2)$ be independent random variables. Then the random variable defined as

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

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

Suppose our null hypothesis is

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

Let $n_j$ be the number of measurements taken from the $j$-th group, and suppose $N = \sum_{j=1}^{k} n_j$. Next, define the sample averages

$$\overline{X}^j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^j$$

$$\overline{X} = \frac{1}{N} \cdot \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_i^j.$$

The difference between these two quantities is that $\overline{X}^j$ is the sample average of data from the $j$-th group, while $\overline{X}$ is the sample average of the data from all the groups combined.
With these defintions, we can now definite the test statistic we will use for the $F$-test. In particular, suppose we let our test statistic be given by

$$\frac{MSG}{MSE},$$

where

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

$$MSE = \frac{1}{N-k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_i^j - \overline{X}^j)^2.$$

Here, $MSG$ refers to the mean square between groups, and $MSE$ 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 the test statistic $MSG/MSE$ is described by an $F$-distribution with $d_1 = k-1$ and $d_2 = N-k$ degrees of freedom. To compute the $p$-value, we can use a table or a calculator to determine

$$p = \mathbb{P}\left(F > \frac{MSG}{MSE}\right),$$

where $F$ is a random variable with $F$-distribution with $d_1 = k-1$ and $d_2 = N-k$ degrees of freedom.

## 2.2 Example: Diet and Lifespan

Q: The following data relate to the ages at death of a certain species of rats that were fed 1 of 3 types of diets. Thirty rats of a type having a short life span of an average of 17.9 months were randomly divided into 3 groups of 10 each. The sample means and variances of ages at death (in months) are:

| | Very Low Calorie | Moderate Calorie | High Calorie |
|---|---|---|---|
| Sample mean | 22.4 | 16.8 | 13.7 |
| Sample variance | 24.0 | 23.2 | 17.1 |

Test the hypothesis, at the 5 percent level of significance, that the mean lifetime of a rat is not affected by its diet.

A: $k = 3$, $n_1, n_2, n_3 = 10$, and $N = n_1 + n_2 + n_3 = 30$.
$H_0 : \mu_1 = \mu_2 = \mu_3$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, where $X^j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ for $j = 1, 2, 3$
Since $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$, for $i = 1, 2, 3$. It follows

$$\begin{aligned}
MSE &= \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2 \\
&= \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i^2 \\
&= \frac{1}{30-3}((10-1)\cdot 24.0 + (10-1)\cdot 23.2 + (10-1)\cdot 17.1) \\
&= 578.7/27 = 21.43
\end{aligned}$$

Also, $\overline{X} = \frac{1}{N}\sum_{i=1}^{k} n_i \overline{X}_i = \frac{1}{30}(10 \cdot 22.4 + 10 \cdot 16.8 + 10 \cdot 13.7) = 17.63$. Then we have

$$\begin{aligned} MSG &= \frac{1}{k-1} \cdot \sum_{i=1}^{m} n_i (\overline{X}_i - \overline{X})^2 \\ &= \frac{1}{3-1} \cdot 10 \cdot [(22.4 - 17.63)^2 + (16.8 - 17.63)^2 + (13.7 - 17.63)^2] \\ &= 388.9/2 = 194.5 \end{aligned}$$

$$TS = \frac{MSG}{MSE} = \frac{194.5}{21.43} = 9.072$$

$$p - value = P(F_{k-1,N-k} > TS) = P(F_{2,27} > 9.072) = 0.00097 < 0.05$$

So we reject $H_0$.

## 2.3 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

$$H_0 : \text{median}(X^1) = \ldots = \text{median}(X^k) \text{ and } f_{X^j}(u) = f_{X^q}(u) \text{ for } j \neq q.$$

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 $N(= \sum_{j=1}^{k} n_j)$ is assigned to each data point in the single list. Next, the test statistic

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

where $r_i^j$ is the rank of the $i$-th data point from the $j$-th group, and

$$\begin{aligned} \overline{r}^j &= \frac{1}{n_j} \cdot \sum_{i=1}^{n_j} r_i^j \\ \overline{r} &= \frac{1}{N} \cdot \sum_{j=1}^{k} \sum_{i=1}^{n_j} r_i^j = (N+1)/2. \end{aligned}$$

The quantity $\overline{r}^j$ is the sample average of ranks of the $j$-th group, and $\overline{r}$ is the sample average of ranks of all the groups combined; however, here we can pre-compute the sample average of ranks of all the groups combined because it is known that the ranks consecutively range from $1$ to $N$.

The intuition of this test is that the test statistic $K$ 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

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

which approximately looks like a $\chi^2(k-1)$ 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 $\alpha$, 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 $\gamma$ should not depend upon the value $p$ otherwise the derivation above may not hold. In other words, $\gamma$ needs to be selected prior to conducting the tests.