Multiple *t* tests or ANOVA (analysis of variance)?

Nural Bekiroğlu

Turkish Respiratory Journal, 2001;2 (1):21-22

In the last issue, we have started to discuss *statistical errors*. Another most common error which is done in statistics, is realised when means of more than two groups are compared.

If the investigator compares set of measurements taken from two groups to decide whether the group means differ, the interpretation of the *t* test result becomes easy. The problem arises when the investigator wants to compare the means of several groups as in the case of the study by Whole et al (1975). They studied the physiological properties of the lungs in the patients who received therapeutic bilateral pulmonary irradiation in early childhood. Their subjects consisted of three groups of children. Group 1 received a single course of bilateral pulmonary irradiation, Group 2 had received additional pulmonary radiotherapy or thoracic surgery or both, Group 3 received no irradiation directed primarily to the lung. Vital capacity values, expressed as percentages of predicted values based on standing height for subjects in the three groups, are measured. If the investigator wants to test the null hypothesis of no difference of the mean of vital capacity values among the three groups, he/she might examine each pair of groups which would involve three tests by using multiple *t* test such as:

Group 1 versus Group 2
Group 1 versus Group 3
Group 2 versus Group 3

(Keeping in mind that any experiment with *k groups* has *k (k-1)/2* different pairs available for testing.)

Correspondence: *Nural Bekiroğlu*
Marmara Üniversitesi Hastanesi, Altunizade, İstanbul, Türkiye

Suppose that if the groups share a common population mean and when pairwise comparisons are performed among three groups separetely; for instance: group 1 versus group 2 or group 2 versus group 3; and tested whether their groups means are equal by *t* test. This must be a possible solution but not an appropriate one: one reason is that if you have more than 3 groups for instance 5 groups to be compared, there will be 10 pairwise tests to be performed so a lot of work and tests to be done. Another reason is that even if the groups are tested by pairs, generalisation can not be drawn for the whole population because of the lack of independence. Therefore group number becomes important. For instance, when there are three groups, 3 pairwise comparisons have to be performed.

If group 1 and group 2 are compared to test whether their group means are equal or not and the null hypothesis is rejected, this means that the test result is considered statistically significant at the α-level or significance level referred as p-value. The 5 percent of significance level is generally preferred but other values can also be selected such as 1% or 10%. This small probability (such as 5%) is the probability of the null hypothesis is falsely rejected. When we want to calculate the probability of obtaining statistically significant test results for three independent tests simultaneously (such as group 1 versus group 2, group 1 versus group 3, group 2 versus group 3) we have to multiply the probabilities of each test individually which has 0.05 of being significant under the null hypothesis. In fact, the probability that three test results will be significant is 0.05 x 0.05 x 0.05= 0.00013.

The probability that none of the three tests will be significant is 0.95 x 0.95 x 0.95 = 0.8574, so the probability

that at least one test will be significant is approximately 14 % or about three times 5%.

In general if n tests have to be performed, the probability of finding at least one wrongly significant independent result can be calculated as follows:

Probability of at least one wrong test result: 1 – (1-)n.

As the number of independent tests increases, the probability of at least one wrong test result or false-positive result also increases and becomes much larger than 0.05.

The formula above assumes that the tests are statistically independent. In this situation there would be no relation between any of the differences between pairs of group means. But when the means of several group are compared, obviously the difference between the means of group 1 and group 2 is related to the difference between the means of group 1 and group 3 and to the difference between the means of group 2 and group 3. Because of this reason, when several groups means are compared and no real population differences exists, multiple *t* tests cause a misleading increase in the probability of finding at least one significant test result. To avoid this problem the analysis of variance (ANOVA) test is recommended.

Analysis of variance enables an extrapolation of the *t* test results of two groups to three or more groups. F-statistic will be calculated for analysis of variance (ANOVA) to test whether group population means are all equal or not. When F-statistic is found significant, we may conclude that at least one of the population means of the groups differs from the others but ANOVA does not tell us which groups are different from which others. If this is the case, a multiple-comparison analysis by pairwise group comparison will be an appropriate answer to this question.

There are several methods for multiple comparison analysis such as Bonferroni Method, Scheffé Method, Tukey Method, Newman-Keuls Method, Duncan Method etc.. Each method assumes that the data are normally distributed and that the population (but unknown) variance within each group is the same.

Many statistical computing packages (such as SPSS, SAS etc.) provide these multiple comparison methods but a biostatistician' consultancy will be very helpfull for choosing the most appropriate multiple comparison method.