Open lecture on statistics

ISSN 2234-7658 (print) / ISSN 2234-7666 (online)
http://dx.doi.org/10.5395/rde.2015.40.2.172

RDE
Restorative Dentistry & Endodontics

Statistical notes for clinical researchers: *post-hoc* multiple comparisons

Hae-Young Kim*

Department of Health Policy and Management, College of Health Science, and Department of Public Health Sciences, Graduate School, Korea University, Seoul, Korea

*Correspondence to
Hae-Young Kim, DDS, PhD.
Associate Professor, Department of Health Policy and Management, College of Health Science, and Department of Public Health Sciences, Graduate School, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, Korea 136-701
TEL, +82-2-3290-5667; FAX, +82-2-940-2879; E-mail, kimhaey@korea.ac.kr

For comparison of three or more group means we apply the analysis of variance (ANOVA) method to decide if all means are equal or there is at least one mean which is different from others. If we get a significant result, we can conclude a global decision that there is difference in group means. However then we need to know what specific pairs of group means show differences and what pairs do not. The procedure is performed by *post-hoc* multiple comparison procedures.

Multiple comparisons and type I error (α error)

Multiple comparisons are procedures of comparing many group means simultaneously. For an example, when we are interested in comparing means of A, B and C groups, we may consider performing a set of three comparisons as following:
Hypothesis 1: mean values of group A and group B are equal (comparison of A and B).
Hypothesis 2: mean values of group A and group C are equal (comparison of A and C).
Hypothesis 3: mean values of group B and group C are equal (comparison of B and C).

The set of comparisons is referred as a 'family of test'. The multiple comparison procedure that tests a set of hypotheses at the same time is also called a 'simultaneous test'. The most important issue in multiple comparisons is the control of type I error. Type I error is defined as the probability of committing error that a true null hypothesis is rejected. We call the type I error as α error. An α error level of 0.05 is frequently used for assessing a hypothesis. The overall error level for the family of tests is different from the α error level for a comparison (Table 1).

Table 1. Type I error rates according to the range of comparison

Category	Definition
Type I error rate per comparison (α_{PC})	Probability of incorrect rejection of true null hypothesis when a comparison is tested
Family-wise type I error rate (α_{FW})	The probability of type I error occurs when a set (family) of tests is performed

If the same α error level is adopted for each comparison (α_{PC}) in multiple k comparisons, the overall α error level for the family of tests (α_{FW}) is calculated as following procedure:

a) Probability of **no** α error for a comparison = 1 - (probability of α error per comparison [α_{PC}]
b) Probability of **no** α error for overall family of **k** tests = $(1 - \alpha_{PC}) \times (1 - \alpha_{PC}) \times \cdots \times (1 - \alpha_{PC}) = (1 - \alpha_{PC})^k$, for comparisons **independent** of each other.
c) Probability of α error for overall family of k independent tests (α_{FW}) = $1 - (1 - \alpha_{PC})^k$

For the example of family of independent three tests, if we set α_{PC} at the conventional α error level 0.05, α_{FW} is obtained as $1 - (1 - 0.05)^3 = 1 - 0.8574 = 0.1426$. The familywise α level is not only greater than the α_{PC} but also greater than an acceptable α error level. If we want to control the below 5 percent (0.05), we need to reduce α_{PC} to a certain degree. For example if we set α_{PC} at 0.01, the α_{FW} is calculated as $1 - (1 - 0.01)^3 = 1 - 0.9703 = 0.0297$, which is smaller than 0.05.

Developed various multiple comparison methods

Many statisticians have devised various multiple comparison methods to correct the α_{FW} within an acceptable level of 0.05. The methods can be categorized into three types according to the size and nature of family of tests, such as restricted sets of contrasts, pairwise comparisons, and *post-hoc* error correction. Table 2 shows the category and characteristics of various multiple comparison tests.

Table 2. Category and characteristics of various multiple comparison tests

Category	Usage	Method	Step	Procedure
Restricted set of contrasts	Appropriate for relatively small families of tests	Bonferroni	Single-step	Adjust $\alpha_{PC} = \alpha_{FW} / k$
		Holm-Bonferroni	Step-down	Repeated tests according to the *p*-values of comparisons and assess with increased α error level
		Shaffer's modified sequentially rejective Bonferroni	Step-down	Modification of Holm-Bonferroni procedure. More powerful.
		Šidák-Bonferroni	Single-step	Adjust $\alpha_{PC} = 1 - (1 - \alpha_{FW})^{\frac{1}{k}}$
		Šidák-Holm	Step-down	Step-down procedure of Šidák-Bonferroni
		Dunnett's test	Single-step	Interested in comparison with control and other experimental groups
Pairwise comparison	Perform all pairwise comparisons	Tukey's honestly significance difference (HSD)	Single-step	Based on studentized range statistics (q statistics). For unequal group size Tukey-Kramer method is applied.
		Tukey's b	Step-down	Step-down procedure of Tukey method. Widely significance difference (WSD) method
		Student-Newman-Keuls (SNK) procedure	Step-down	Step-down procedure of modified Tukey method. Lower critical values are applied as steps repeated. Very liberal.
		Duncan's multiple range test	Step-down	Modified SNK method. Adjust significance level as steps repeated.
		Hochberg's GF2	Single-step	Similar to Tukey method but critical values are based on the studentized maximum modulus distribution. appropriate for balanced or unbalanced design
		Gabriel	Single-step	Equivalent to GF2 test for balanced one-way ANOVA. For unbalanced design, less conservative than GF2
		Ryan-Einot-Gabriel-Walsch (REGW)	Step-down	Modification of the SNK procedure (more strict control of family-wise α error) REGWF: based on F statistics (available only in SPSS) REGWQ: based on *q* statistics (maximum range distribution)
Post-hoc error correction	For a completely *post-hoc* analysis	Scheffé	Single-step	Complex comparison Most stringent error control
No control of error	Inappropriate for control of family-wise error	LSD (Least Significant Difference)	Single-step	No control of family-wise error

α_{PC}, α error level per comparison; α_{FW}, family-wise α error level; k, number of comparisons.

1. Restricted sets of contrasts

The multiple comparison methods in ‘Restricted sets of contrasts’ are appropriate for relatively small families of tests composed of less than ten tests (or contrasts) approximately. The results by the methods in the category can be somewhat conservative when applied for a large number of tests. Therefore before the test, the contrasts of specific interest should be chosen such as a set of planned comparisons or comparisons between a control group and other experimental groups.

Before introducing each specific method a brief overview of related terms in statistics may help.

- Liberal/conservative: ‘Liberal’ refers to a tendency that rejection of null hypothesis is relatively easy. A liberal test has a large power in accepting true alternative hypothesis. Contrarily, ‘conservative’ refers a relative difficulty in rejecting null hypothesis and possession of small power.
- Balanced design/unbalanced design: Balanced design refers a condition that all the compared groups have equal sample sizes; unbalanced design refers that the groups have unequal sample sizes.
- Single-step/step-down methods: A single step test refers a test which is implemented by a single test procedure; a step-down method refers a test which is performed according to repeated sequential steps. Generally while a single step test provides a confidence interval, a step-down method does not.
- Common step of step-down methods: At first step, all k means are tested at a α_{FW} level considering comparison of k means. If the result is significant, then the following step starts; if insignificant, it stops. Next, each subset of k - 1 means is tested at an increased α_{FW} level considering comparison of k - 1 means. Continue in this manner until no subsets remain to be tested.

1.1. Single step methods

1.1.1. The Bonferroni procedure

The Bonferroni correction of α error is a completely general method which is widely applicable to any sort of statistical procedures other than multiple comparisons following ANOVA. The α error for overall family of k independent tests (α_{FW}) ‘$1 - (1 - \alpha_{PC})^k$’ is the largest value among α_{FW} of any set of tests including both independent and correlated tests. After some algebra the Bonferroni inequality for any set of tests is expressed as $\alpha_{FW} < k\alpha_{PC}$. Therefore, to control the α_{FW} to be smaller than 0.05, we apply α error level for each comparison as $\alpha_{PC} = \alpha_{FW}/k$. For a family of three tests as the example above, we apply $\alpha_{PC} = \frac{0.005}{3} \cong 0.0167$ to control α_{FW} as 0.05. The limitation of the Bonferroni procedure is that the result tends to be too conservative and do not have enough power when the set of tests is large.

1.1.2. The Šidák-Bonferroni procedure

The Šidák-Bonferroni procedure was developed to improve the power of tests because the Bonferroni procedure produces conservative results. The significance level per comparison is applied as $\alpha_{PC} = 1 - (1 - \alpha_{FW})^{\frac{1}{k}}$. For a family of three tests as the example above, we apply $\alpha_{PC} = 1 - (1 - 0.05)^{\frac{1}{3}} \cong 1 - 0.9830 = 0.0170$ to control α_{FW} as 0.05. By applying a slightly higher α_{PC} level, the result produces a less conservative result compared to the Bonferroni procedure. Let’s consider an example of comparing three groups. We got a *p*-value 0.0168 for a comparison of two group means. If we apply the Bonferroni procedure, we conclude that difference between two group means is insignificant because the calculated *p*-value is over the level ($\alpha_{PC} = 0.0167$). However if we apply the Šidák-Bonferroni procedure, we reach the conclusion of a significant difference.

1.1.3. Dunnett’s test

Specifically when one control group is being compared to all other experiment groups, the Dunnett’s test is appropriate. In the situation the Dunnett’s test shows a large power. The standard *t* which is uncorrected is used as a test statistic, and compared with the particular value of $t_{Dunnett}$ that Charles Dunnett devised.

1.2. Step-down procedures

1.2.1. Holm-Bonferroni procedure

A step-down repeated test similar to the Bonferroni procedure is performed according to ordered *p*-value of each comparison. As the step progress, comparisons are assessed with successively increased α error levels. It shows more power compared to the Bonferroni procedure.

1.2.2. Shaffer's modified sequentially rejective Bonferroni procedure

It is a modification of Holm-Bonferroni procedure by partly adopting increased α error levels, having more power compared to the Holm-Bonferroni procedure.

2. Pairwise comparisons

The pairwise comparison is comparing all possible pairs of group means. If we want to compare all possible pairs from k groups, then the total number of comparisons is k(k - 1)/2. Following procedures are appropriate for all pairwise comparison and are expected to obtain reasonable results. Though it is possible to apply Bonferroni correction, overcorrected result is expected.

2.1. Single step methods

2.1.1. Tukey's honestly significant difference (HSD) procedure

Tukey's HSD procedure provides the simplest way to control α_{FW} and is considered as the most preferable method when all pairwise comparisons are performed. The studentized range statistic (q statistic) is used to determine the critical values based on number of groups and number of observations in a group. As Tukey's HSD procedure assumes equal size of all compared groups, a modified Tukey-Kramer method can be applied for comparisons of unequal-sized groups.

2.2. Step-down procedure

2.2.1. Student-Newman-Keuls (SNK) procedure

The Student-Newman-Keuls (SNK) procedure is a step-down procedure which constructs equivalent subset similar to Tukey's procedure. The SNK procedure is following a very complex process. Though it shows an increased power, it often comes with an increased family-wise error level and may result in a too liberal tendency.

2.2.2. Duncan's multiple range test

The Duncan's multiple range test is performed using steps similar to SNK procedure. Changed α error levels are applied following the step-down procedure. It shows more liberal tendency than the SNK procedure. Generally Duncan's multiple range test is not recommended when sample sizes are unequal because of the liberal tendency.

2.2.3. Ryan-Einot-Gabriel-Walsch (REGW) procedure

The REGW procedure is a modification of SNK procedure by introducing more strict control of family-wise α error. The REGW procedure is considered to be recommendable because it shows not only good power and but also tight error control. The REGWF uses F statistic and REGWQ uses the studentized range statistic (q statistic).

3. *Post-hoc* error correction

The procedure in this category is performed as a completely *post-hoc* analysis after all planned comparisons are assessed. The procedure explores all possible complex relationships and applies with the most stringent error control.

3.1 Scheffé's procedure

The Scheffé's procedure comprises all possible contrasts not only paired comparisons. Its advantage is that it covers a broad range of complex tests including *post-hoc* relationships among many groups. The procedure tends to be too conservative and power is less than other methods. Generally Scheffé's procedure is not recommended when only pairwise comparisons are of interest.

4. No control of family-wise α error level

4.1. Least Significant Difference (LSD) test

The LSD method does not control family-wise α error level. Therefore it is inappropriate for multiple comparison procedure where the control of family-wise α error level is necessary.

Summary of multiple comparison methods

In the choice of multiple comparison methods, it is important to consider the exact situation. The standard of choice is the ability to control family-wise α error level and the degree of power detecting significant difference.

1. For usual *post-hoc* pairwise comparisons, Tukey's HSD procedure or REGWQ may be preferable.
2. For comparisons of small number of group means or preplanned comparisons of selected groups, the Bonferroni procedure or Šidák-Bonferroni procedure may be preferable.
3. When a control group is compared with other experimental groups, Dunnett's test may be of choice.
4. If interested in a broad range of complex tests, Scheffé's procedure may be appropriate.

Note: For practical convenience, most statistical packages show adjusted *p*-values which are comparable with a conventional α error level instead of the reduced α_{PC}. Clinical researchers may simply compare *p*-values provided by statistical packages with conventional α error level such as 0.05 and may make decisions comfortably.

Reference

1. Keppel G, Wickens TD. Design and analysis: a researcher's handbook. 4th ed. New Jersey: Pearson Education Inc.; 2004. p111-130.