Post-hoc and multiple comparison test – An overview with SAS and R Statistical Package

Editor, International Journal of Statistics and Medical Informatics (ISJMI)

Abstract

Analysis of Variance (ANOVA) is a basic but most important tool in Statistics. The simplest form is one way ANOVA wherein equivalence of treatment means are tested. If the means are not equal then the next step is to check which means are different from each other. Post-Hoc and multiple comparison tests are used to identify which pairs of treatment means differ. This paper starts with the overview of Post-Hoc and Multiple Comparison test and discusses the various Post-hoc multiple comparison tests, its usability, positives and limitations. The paper also provides the Statistical Analysis System (SAS) and R Statistical Package codes to carry out the various Post-hoc and multiple comparison tests

Keywords: Analysis of Variance; Post-hoc; Post hoc; multiple comparison; ANOVA; SAS; R package

Introduction

When a researcher wants to test the difference between three or more drug's effect on controlling serum cholesterol, one way Analysis of Variance (ANOVA) method can be used to find out whether mean cholesterol level between the groups are statistically different or not. If there is a difference between the drugs effects found on controlling the serum cholesterol then Post-hoc and multiple comparison tests [1, 2, 3] can used to find out which pair of drugs differ from each other.

Normally a student t-test or equivalent non parametric test such as Mann Whitney U test can be used to test whether the two means differ or not. The problem with the t-test or its non-parametric equivalent tests is that it increases the overall Type I error or family wise error rate.

Type I error and Family Wise Error Rate (FWER)

The type 1 error for single test (error wrongly rejecting the null hypothesis when it is actually true) is denoted by α.

Decision from the Test	Null Hypothesis True Status : True	Null Hypothesis True Status : False
Decision from the test : Reject Null Hypothesis based on the sample data	False Positive (α-Type I error)	True Positive (1- α)
Decision from the test : Accept Null Hypothesis based on the sample data	True Negative (1-β)	False Negative (β- Type II error)

Table: 1 – Type I and Type II error

If n multiple tests are carried out then the cumulative or Family Wise Error Rate is calculated as below

$\alpha_n = 1 - (1 - \alpha_i)^n$ where n is the number of comparison being tested and i is the i^{th} comparison

Post Hoc and Multiple Comparison Tests

Post Hoc tests are used to compare the pair of treatment means while controlling the Family Wise Error Rate. These tests are conducted normally after the one way ANOVA [4, 5] returns significant results.

The following sections discusses various multiple comparison tests [6] available with its usefulness and limitations

1. **Bonferroni-adjusted multiple t-tests(Dunn)** [1,2]
   Bonferroni adjusted multiple t-tests (Dunn) is easy to compute and flexible to use for any multiple comparison while controlling the FWER. It does not require an ANOVA to be significant as it falls under planned comparison procedure.
   Positives
   a. It is easy to compute
   b. It controls the FWER
   c. It does not require ANOVA to be significant
   Limitations
   a. Its lacks power due to the fact that it assumes null hypothesis is true for all the tests in consideration
2. **Sidak test** [1,2]
   Sidak test is having slightly higher power than Bonferroni test while retaining the FWER
   Positives
   a. It is easy to compute
   b. It controls the FWER
   c. Slightly powerful than Bonferroni
   Limitations
   a. Lacks power
3. **Dunnett's test**[1,2]
   Dunnett's test tests only compare the control group with the other groups. . In each pair wise comparisons control group will be present. It does not compare the other groups with each other
   Positives
   a. It is exact procedure to compare a control with the other groups in consideration
   Limitations
   a. Limited application due to the fact that it is useful when one compares the control with the other groups
   b. Equal variance assumption to be met

4. **Tukey honestly significant difference (HSD) test** [8]
   Tukey's test works on the Studentized range statistic called Q statistics which calculates the critical value based on the number of groups and number of sample observation in the

group. Tukey's test assumes the sample observations being tested are independent within and between the groups, group means are normally distributed and assumes equal variance among the group.

Positives

a. It maintains the alpha level at the desired range when the three assumptions are met
b. It is useful when sample sizes are not equal among groups and
c. All pairwise comparisons are carried out

Limitations

a. Powerful
b. It assumes equal variances for the groups which may not be the case always

5. **Games and Howell's modification of Tukey's HSD** [9]

It is modification of Tukey's HSD test and used when the unequal variance assumption is violated

Positives

a. It is useful when unequal variances assumption is violated while using the Tukey's HSD test

Limitations

a. It is less conservative when the sample sizes of the groups are small.

6. **Tukey's wholly significant difference (WSD) test** [1,2]

It is modification of Tukey's HSD test and used when the unequal variance assumption is violated

Positives

a. It is useful when unequal variances assumption is violated while using the Tukey's HSD test

Limitations

a. It is less conservative when the sample sizes of the groups are small.

7. **Newman-Keuls test(Student-Newman-Keuls- SNK)** [10]

It is one of the step down procedure where in the difference between the largest and smallest means are compared first if it is significant continue the next set of pairs (second largest vs smallest or second smallest vs largest) or stop if the pair is not significant. This test is continued till a non-significant pair comparison is reached

Positives

a. This test is useful when the number of pairwise comparisons are more
b. Liberal than Dunn's Test

Limitations

a. FWER is not controlled

8. **Ryan Einot Gabriel Welch q test (REGWQ)** [1,11]

It is one of the step down procedure where in the difference between the largest and smallest means are compared first if it is significant continue the next set of pairs (second largest vs smallest or second smallest vs largest) or stop if the pair is not significant. This test is continued till a non-significant pair comparison is reached

Positives

a. FWER is controlled

b. Liberal and powerful than Tukey's Test

Limitations

a. Not useful when the sample size between the groups are different
b. Less powerful than REGWF test

9. **Ryan Einot Gabriel Welch F test (REGWF))** [1,11]

This test is based on F statistic rather than the q statistic

Positives

a. FWER is controlled
b. Powerful than REGWQ test

Limitations

a. Not useful when the sample size between the groups are different

10. **The Shaffer-Ryan test)** [1,11]

It is a modified version of REGWF test which

Positives

c. FWER is controlled
d. Powerful than REGWQ and REGWF test

Limitations

b. Not useful when the sample size between the groups are different

11. **The Least Significant Difference test (LSD)/ Fisher's LSD** [12]

It is the simplest of the entire multiple comparison test and it controls the FWER only when 3 means are tested.

Positives

a. Most Powerful test

Limitations

a. Very poor in controlling the FWER

12. **The Fisher-Hayter test** [13]

It is a modification of Fishers LSD test

Positives

a. Easy to use
b. Controls the FWER

Limitations

a. Useful only for pairwise contrasts
b. Less Powerful when number of comparisons are less

13. **Waller-Duncan test** [1,2]

It is different from the other multiple comparison tests as it uses Bayesian approach wherein it minimizes the overall loss function which is a sum of loss functions for all comparisons

14. **Dunnett's T3** [1,2]

This test is used to test the pairwise comparison, groups are having unequal variance and group sample size is small.

Positives

a. Controls FWER

b. It is useful when variances of the groups are unequal

c. It is useful when the sample size is small

Limitations

a. FWER may exceed the desired level when the variances are equal

15. **Tamhane' T2** [1,2,14]

This test is used to test the pairwise comparison and groups are having unequal variance

Positives

a. Controls FWER

b. It is useful when variances of the groups are unequal

Limitations

a. FWER may exceed the desired level when the variances are equal

16. **Howell and Dunnett's C** [1,2]

This test is used to test the pairwise comparison; groups are having unequal variance. This test is useful when the sample size is large

Positives

a. Controls FWER

b. It is useful when variances of the groups are unequal

c. It is useful when the sample size is large

Limitations

b. FWER may exceed the desired level when the variances are equal

17. **The Tukey-Kramer test** [13]

It is a modified version of Tukey's test when the groups are having unequal sample size

Positives

a. When the sample size is unequal

Limitations

a. Useful only when fixed number of comparisons are used

18. **The Miller-Winer test** [1,2]

This test is used when the sample size of the groups are unequal.

Positives

a. When the sample size is unequal

19. **Hochberg's GT2 test** [1,2]

This test is used when the sample size of the groups are unequal.

Positives

a. When the sample size is unequal

20. **Gabriel test** [15]

This test is used when the sample size of the groups are unequal.

Positives

b. When the sample size is unequal

21. **The Scheffe test** [16]

Scheffe test is a flexible and conservative test and is useful for both simple and complex comparisons.

Positives

a. Controls FWER

b. Useful when more number of comparison to be used
c. Useful for both equal and unequal sample sizes

Limitations

a. It has less power

22. The Duncan's Multiple Range test [17]

It is a modified version of Student Newman Keuls method

Positives

d. It is powerful

Limitations

b. It does not control the FWER

Example using SAS® software [7]

```
data cholestral;
input treatmentgroup serumcholestrol;
datalines;
2 253
2 253
2 324
2 303
2 247
3 331
2 332
1 319
3 202
2 211
4 230
3 202
3 216
4 220
3 305
3 223
2 229
3 213
3 292
3 211
3 293
1 330
2 264
1 239
4 295
4 277
3 292
2 269
```

4 338
1 299
3 246
3 273
1 245
3 313
2 309
3 212
3 240
2 207
1 291
3 302
2 211
2 273
3 334
4 269
1 234
4 217
4 336
2 253
3 207
3 228
4 252
2 293
4 257
1 240
2 307
3 280
2 294
3 337
4 350
4 274
4 302
4 283
4 320
3 288
4 212
1 248
4 236
4 340
3 310
2 279
2 310
4 331
1 288

3 222
2 300
2 206
1 204
2 268
2 237
2 226
3 308
;

```
proc glm;
class treatmentgroup;
model serumcholestrol = treatmentgroup;
means treatmentgroup/Lsd tukey bon Duncan Dunnett Gabriel regwq scheffe
sidak waller;
run;
```

Example using R Statistical Package Code

>treatmentgroup<-
c(3,1,3,4,2,1,4,1,4,3,2,1,3,1,4,4,3,2,2,1,3,4,1,1,4,2,4,3,1,2,1,2,3,4,4,2,1,1,4,3,1,2,4,1,3,2,3,1,2,3,1,4,1,2,3,2,2,3,3,4,2,4,4,3,4,4,1,2,1,4,2,3,3,3,2,4,2,1,1,3,3)
>serumcholestrol<-
c(245,198,122,308,261,282,221,191,191,209,282,274,285,193,265,273,229,207,257,375,273,269,204,204,229,281,289,309,290,270,217,190,286,241,264,234,221,200,227,259,226,237,264,218,255,262228,320,247,230,251,386,383,322,369,203,213,222,283,227,260,232,209,272,213,244,224,287,231,230,290,269,237,257,295,429,225,262,224,235,287,293)
>tapply(serumcholestrol, treatmentgroup, mean)
pairwise.t.test(serumcholestrol, treatmentgroup, p.adj = "none")
pairwise.t.test(serumcholestrol, treatmentgroup, p.adj = "bonferroni")
pairwise.t.test(serumcholestrol, treatmentgroup, p.adj = "holm")
pairwise.t.test(serumcholestrol, treatmentgroup, p.adj = "hochberg")
pairwise.t.test(serumcholestrol, treatmentgroup, p.adj = "hommel")
pairwise.t.test(serumcholestrol, treatmentgroup, p.adj = "BH")
pairwise.t.test(serumcholestrol, treatmentgroup, p.adj = "fdr")
pairwise.t.test(serumcholestrol, treatmentgroup, p.adj = "BY")

Conclusion

The paper discussed the various post hoc multiple comparison tests, its usefulness and limitation and also provided the SAS and R statistical package codes with the example dataset

References

[1]. Toothaker, L. E. (1993). Multiple comparison procedures (No. 89). Sage.

[2]. Saville, D. J. (1990). Multiple comparison procedures: the practical solution. *The American Statistician*, *44*(2), 174-180.

[3]. Kim, H. Y. (2015). Statistical notes for clinical researchers: post-hoc multiple comparisons. Restorative dentistry & endodontics, 40(2), 172-176.

[4]. Ruxton, G. D., & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. Behavioral Ecology, 19(3), 690-693.

[5]. Cabral, H. J. (2008). Multiple comparisons procedures. Circulation, 117(5), 698-701.

[6]. Brown, A. M. (2005). A new software for carrying out one-way ANOVA post hoc tests. Computer methods and programs in biomedicine, 79(1), 89-95.

[7]. www.vinaitheerthan.com

[8]. Abdi, H., & Williams, L. J. (2010). Tukey's honestly significant difference (HSD) test. Encyclopedia of Research Design. Thousand Oaks, CA: Sage, 1-5.

[9]. Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: a Monte Carlo study. Journal of Educational and Behavioral Statistics, 1(2), 113-125.

[10]. Abdi, H., & Williams, L. J. (2010). Newman-Keuls test and Tukey test. Encyclopedia of Research Design. Thousand Oaks, CA: Sage, 1-11.

[11]. Ryan, T. H. (1960). Significance tests for multiple comparisons of proportions, variances, and other statistics. Psychological bulletin, 57(4), 318.

[12]. Williams, L. J., & Abdi, H. (2010). Fisher's least significant difference (LSD) test. Encyclopedia of research design, 1-5.

[13]. Richter, S. J., & McCann, M. H. (2012). Using the Tukey–Kramer omnibus test in the Hayter–Fisher procedure. *British Journal of Mathematical and Statistical Psychology*, *65*(3), 499-510.

[14]. Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association*, *74*(366a), 471-480.

[15]. Gabriel, K. R. (1969). Simultaneous test procedures--some theory of multiple comparisons. *The Annals of Mathematical Statistics*, 224-250.

[16]. Scheffe, H. (1999). *The analysis of variance* (Vol. 72). John Wiley & Sons.

[17]. Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrics*, *11*(1), 1-42.